Published in final edited form as:

Comput Toxicol. 2018 August ; 7: 46–57. doi:10.1016/j.comtox.2018.06.003.

Novel application of normalized pointwise mutual information (NPMI) to mine biomedical literature for gene sets associated with disease: use case in breast carcinogenesis

Sean M. Watford^{1,2}, Rachel G. Grashow^{3,4}, Vanessa Y. De La Rosa^{3,5}, Ruthann A. Rudel³, Katie Paul Friedman⁷, Matthew T. Martin^{6,7}

¹ORAU, contractor to U.S. Environmental Protection Agency through the National Student Services Contract, Oak Ridge, TN.

²Department of Environmental Sciences and Engineering, Gillings School of Global Public Health, UNC-Chapel Hill, Chapel Hill, North Carolina, United States.

³Silent Spring Institute, Newton, MA.

⁴Department of Environmental Health, Harvard T.H. Chan School of Public Health, Boston, MA

⁵Social Science Environmental Health Research Institute, Northeastern University, Boston, MA

⁶U.S. Environmental Protection Agency, National Center for Computational Toxicology, Research Triangle Park, NC, USA.

⁷Currently at Pfizer Worldwide Research & Development, Groton, CT, USA.

Abstract

Advances in technology within biomedical sciences have led to an inundation of data across many fields, raising new challenges in how best to integrate and analyze these resources. For example, rapid chemical screening programs like the US Environmental Protection Agency's ToxCast and the collaborative effort, Tox21, have produced massive amounts of information on putative chemical mechanisms where assay targets are identified as genes; however, systematically linking these hypothesized mechanisms with *in vivo* toxicity endpoints like disease outcomes remains problematic. Herein we present a novel use of normalized pointwise mutual information (NPMI) to mine biomedical literature for gene associations with biological concepts as represented by Medical Subject Headings (MeSH terms) in PubMed. Resources that tag genes to articles were

Corresponding author: Matthew Martin, Previously at US EPA, Now at Pfizer Worldwide Research and Development, 445 Easter Point Rd, MS 8274-1224, Groton, CT 06340, Phone: 860-441-4510, Matthew.Martin@pfizer.com.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Disclaimer: The United States Environmental Protection Agency (U.S. EPA) through its Office of Research and Development has subjected this article to Agency administrative review and approved it for publication. Mention of trade names or commercial products does not constitute endorsement for use. The views expressed in this article are those of the authors and do not necessarily represent the views or policies of the US EPA.

Conflicts of Interest

The authors declare no conflicts of interest.

integrated, then cross-species orthologs were identified using UniRef50 clusters. MeSH term frequency was normalized to reflect the MeSH tree structure, and then the resulting GeneID-MeSH associations were ranked using NPMI. The resulting network, called Entity MeSH Co-occurrence Network (EMCON), is a scalable resource for the identification and ranking of genes for a given topic of interest. The utility of EMCON was evaluated with the use case of breast carcinogenesis. Topics relevant to breast carcinogenesis were used to query EMCON and retrieve genes important to each topic. A breast cancer gene set was compiled through expert literature review (ELR) to assess performance of the search results. We found that the results from EMCON ranked the breast cancer genes from ELR higher than randomly selected genes with a recall of 0.98. Precision of the top five genes for selected topics was calculated as 0.87. This work demonstrates that EMCON can be used to link *in vitro* results to possible biological outcomes, thus aiding in generation of testable hypotheses for furthering understanding of biological function and the contribution of chemical exposures to disease.

Keywords

Biomedical Literature; Data Integration; Genes; Breast Carcinogenesis; Chemical Exposures; Literature Mining

1 Introduction

Rapid technological advances in biomedical and life sciences have led to an inundation of heterogeneous information across these fields. For instance, high-throughput technologies like RNA-Seq and other transcriptomics methods generate large amounts of gene-related data such that underlying biological processes can be illuminated through pathway enrichment or association [1, 2]. Linking these functional genomic data to well-characterized ubiquitous diseases such as breast cancer [3] could provide opportunities to derive additional etiological insight, generate new hypotheses, identify critical genes and pathways, and develop novel therapeutics. However, while the current volume of genetic, experimental, toxicological and other data presents an incredible opportunity for biomedical and basic science to make great knowledge gains, many challenges remain in understanding how this information can be best integrated and queried to produce valuable insight.

Currently, there is no research precedent on how to link genetic and toxicological data to complex disease phenotypes. For example, given that breast cancer will affect one in eight U.S. women and that susceptibility is shaped by both genetic and environmental factors [4–9], it is worthwhile to query publicly available data resources to better understand how risk factors like chemical exposures initiate biological changes to increase disease risk [10–12]. Such an approach represents a departure from conventional toxicological strategies, in which a single chemical exposure is investigated as the driver of adverse effects, rather than considering other components of risk, e.g. genetic and lifestyle factors [13]. As an alternative, the strategy outlined in this study aligns with the Adverse Outcome Pathway (AOP) conceptual framework that focuses on aggregating information on perturbed systems across levels of biological organization [14, 15]. However, development of AOPs faces the same challenges in linking molecular initiating events (MIE), subsequent key events (KE),

and adverse outcomes (AOs) in that the most relevant MIEs of KEs for a given AO may not be known, and it may be difficult to link specific risk factors such as chemical exposures to the AO of concern. With respect to breast cancer, it is possible that many chemicals or mixtures contribute to risk through many mechanisms; therefore, it may be more pertinent to work backwards from the AO to better understand the etiology and more effectively identify KEs and MIEs that may lead to increased breast cancer risk [16]. However, without a comprehensive data science resource that can integrate gene identifiers or related information on early KEs in toxicity or disease with AOs, the considerable amount of information already available from academic, public, and private sector research may not be fully leveraged for hypothesis generation regarding mechanisms of toxicity or disease. The goal of the work presented herein is to provide just this kind of resource that can provide a putative, ranked linkage between an AO of concern and a given “entity,” i.e., a gene, biological process, or chemical; the result of using this new tool is a ranked list of potentially relevant entities that can be evaluated in follow-up screening, representing a quantitative approach to literature review and hypothesis generation.

Currently there are several high-profile, publicly-available efforts in toxicology developed to explore how chemicals perturb biological systems, including the US Environmental Protection Agency’s Toxicity Forecaster (ToxCast) [17] and the larger, interagency collaboration Tox21 [18]. The high-throughput bioactivity results from these research programs have been used for chemical screening efforts of regulatory importance, like the Endocrine Disruptor Screening Program (EDSP) [19]. These data have also been useful in research and development of putative AOPs, wherein the chemical-target interactions from high-throughput screening can be used as MIEs [20]. However, even with this considerable amount of information, linking high-throughput screening data to AOs like diseases can be challenging without integration with information at various levels of biological complexity that consider toxicity. Another effort in linking chemical exposures to disease is the Comparative Toxicogenomics Database (CTD). In CTD, chemical-gene interactions are manually curated from published articles and then connected to diseases via inference [21, 22]. The chemical-disease inference is based on overlapping genes for a given chemical-gene pair where the disease-related genes are either manually curated from articles or pulled in from another publicly available resource, Online Mendelian Inheritance in Man (OMIM) [23]. While these disease sources are helpful for exploring gene-disease associations for inherited variants, they do not consider genes involved with the initiation and progression of a disease from non-inherited (i.e. environmental) risk factors. Finally, massive data generation efforts for toxicogenomics are ongoing with applications like the Connectivity Map (CMap) [24] and the S1500+ from Tox21 [25, 26]; in these efforts, analysis of gene expression changes resulting from chemical exposures are being used for drug discovery and repositioning as well as understanding chemical mechanisms of toxicity. With these large data generation activities underway, it is more important than ever that toxicologists have access to a tool that can enable ranked associations between gene identifiers or early KEs and possible AOs; such a tool would require integration of multiple types and sources of data or information.

The most substantial source of biological and biomedical information is PubMed, a freely available database managed by the National Library of Medicine (NLM) that contains over

27 million scientific articles that are indexed by medical subject headings (MeSH terms) [27, 28]. MeSH terms are arranged in a hierarchical tree with parent-child relationships such that each parent encompasses the concepts of each of its descendants, i.e. child MeSH terms are narrower in scope than their broader-scoped parents. For example, as seen in Figure 1, “Ductal, Carcinoma, Breast” has one immediate parent from two branches: “Breast Neoplasms”. “Carcinoma, Ductal, Breast” is a narrower concept than its parent and other ancestors in the tree. MeSH terms cover topics across all the articles within PubMed, but the most relevant topics to the work presented here are those on diseases, symptoms, processes, and related biological and chemical entities, which are well-represented in MeSH terms. These MeSH terms need to be linked with gene identifiers, which requires some additional consideration as gene identifiers are not automatically tagged to articles in PubMed.

Although some genes can be specifically identified by MeSH terms, not all genes are represented, especially since gene identifiers are species-specific. Systematic approaches for tagging genes that use strategies like named entity recognition (NER) have been implemented [29]. However, despite the most successful efforts [30], no global approach exists where all genes can be systematically identified across all articles in PubMed. In lieu of a global systematic approach, we can rely on numerous manual curation efforts with publicly available resources that tag articles with relevant unique gene identifiers (GeneID). Although manual curation efforts are low throughput, the quality of mappings is higher, especially in resources built based on their manual curation efforts, including CTD and Universal Protein Resource (UniProt/SwissProt) [31]. These manually curated resources offer valuable information alone, but also have great potential for discovery if tied together into one larger resource that also includes CTD’s chemical-gene interactions and UniProt’s protein-specific topics.

Here we describe a novel and transferrable methodology, producing a resource known as the Entity MeSH Co-occurrence Network (EMCON), that integrates several resources to develop a network connecting genes to MeSH terms. EMCON uses ranked associations between genes and MeSH terms to produce ranked gene sets for hypothesis generation and testing. The utility of EMCON was demonstrated by evaluating genes putatively linked to breast carcinogenesis, an example highly relevant to public health. Given the breadth of ongoing research on chemicals known to increase the frequency of breast tumors in animals and humans [4–9], important information about breast cancer mechanisms and risk could be uncovered by integrating already existing data housed in resources mentioned above. In this example, MeSH terms were selected that represent important processes in carcinogenesis and, more specifically, breast carcinogenesis. These MeSH terms were used to query EMCON and retrieve a ranked list of genes. Previously, Silent Spring Institute (SSI) created a list of nearly 300 genes as a reference gene set for breast carcinogenesis through expert literature review (ELR) [32]. For the purposes of this study, that reference gene list was used to measure relevance of the EMCON search results. This work demonstrates a novel application of NPMI that critically informs hypothesis generation regarding genes that may be involved in breast carcinogenesis. EMCON may be useful in prioritization and selection of gene sets for transcriptomic experiments and/or articles to be manually reviewed for reference information. The methods described here are transferrable to any disease or AO of interest and could be tailored to myriad biomedical or life science research questions.

2 Methods

The overall workflow detailed in this paper is represented in Figure 2. First, we integrated seven resources, including gene2pubmed [33, 34], Gene Reference into Function (GeneRIF) [34, 35], CTD [21, 36], UniProt/SwissProt [31, 37], Reactome [38, 39], Rat Genome Database (RGD) [40, 41], and Mouse Genome Informatics (MGI) [42, 43], to develop a network of naive GeneID-MeSH associations. Parent MeSH terms are less specific than their child terms, and as such, associations with parent MeSH terms may give less specific insight into the AO (see Figure 1); these parent MeSH terms should be mapped to all articles mapped to their child terms to reflect this relationship. Accordingly, we normalized the MeSH term frequencies to reflect the hierarchy within the MeSH tree so that broader terms appeared in associations more frequently while narrower, more specific terms showed up less often. This resulted in the less specific parent MeSH terms being mapped more promiscuously. Lastly, the GeneID-MeSH associations were ranked using an association measure called normalized pointwise mutual information (NPMI) [44]. NPMI is commonly used in text mining for collocation extraction to identify words that co-occur together more than expected by random chance; the NPMI for any given association is a continuous value between -1 and 1 . An NPMI greater than zero indicates a co-occurrence with greater probability than chance, with increasing significance of the probability as the NPMI value approaches 1 . A positive NPMI does not indicate the direction of association (positive or negative association) between the GeneID and MeSH term. The final resource of ranked GeneID-MeSH associations is represented by EMCON: a scalable, queryable resource for retrieval of a ranked list of genes for a specific topic covered within MeSH terms.

2.1 Integration of biomedical text resources

To build a network relating genes to MeSH terms, we first identified biomedical databases that manually link genes or gene products to relevant articles. These databases included Comparative Toxicogenomics Database (CTD) [21, 36], gene2pubmed [33, 34], Gene Reference into Function (generif) [34, 45], Universal Protein Resource/Swiss-Prot (UniProt) [31, 37], Reactome [38, 39], Rat Genome Database (RGD) [40, 41], Mouse Genome Informatics (MGI) [42, 43]. Each of these resources provide cross-references to Entrez Gene (GeneID) along with a PubMed Identifier (PMID) that uniquely identifies a gene and PubMed article respectively. Entrez Gene is a resource managed by the National Center for Biotechnology Information (NCBI) providing unique identifiers for genes and linking information to genes (biological function, gene products, sequences, etc.) as this type of information is discovered [34]. PubMed is also managed by NCBI and is the largest resource of freely accessible biomedical text with over 27 million citations from a variety of sources including peer-reviewed, biomedical journals. Each of the resources listed above can be integrated into a single resource that links PMIDs with GeneIDs (Figure 2A).

Next GeneIDs were linked to concepts across PubMed via Medical Subject Headings (MeSH terms; Figure 2A). Articles within PubMed are both manually and automatically tagged with MeSH terms, which is a controlled vocabulary of over 27,000 keywords structured in hierarchical trees used to categorize the concepts covered in an article [28]. For example a publication titled “Estrogen receptor variant messenger RNA lacking exon 4 in

estrogen- responsive human breast cancer cell lines“ [46] has been tagged with MeSH term “Breast Neoplasms”, “Receptors, Estrogen”, “RNA, Messenger”, and others. This is exemplified in Figure 3, which shows how GeneIDs are mapped to MeSH terms. This article has also been manually tagged with the gene estrogen receptor alpha (ESR1). Combined with another article [47] that is also manually tagged with ESR1 and a few overlapping MeSH terms, the gene ESR1 now has two articles that support a relationship to the MeSH term “Molecular Sequence Data”, “Amino Acid Sequence”, “Receptors, Estrogen”, “RNA, Messenger”, and “Tumor Cells, Cultured” (Figure 3). As more articles are added to the network, the number of supporting articles for a GeneID-MeSH association grows. Integration of these resources yielded a network of naive GeneID-MeSH associations (Figure 2A).

2.2 Cross-species gene orthologs

Several of the resources above do not exclusively contain information on human genes, but for this work, we are only concerned with human genes. To maximize the number of articles and avoid excluding those that do not have human genes from the network altogether, we identified human orthologous genes. We assumed that the topics from an article about non-human orthologs are relevant to humans. We utilized UniProt Reference Clusters (UniRef), specifically UniRef50, to identify non-human proteins that have a human reference protein with a similar sequence [48]. Proteins were mapped back to the naive GeneID-MeSH network via GeneID cross-references from UniProt/SwissProt. Then all articles tagged with the non-human GeneIDs were mapped back to the reference human GeneID from the corresponding similarity cluster (Figure 2B).

2.3 Medical Subject Heading (MeSH term) Frequency Normalization

At the top of the MeSH hierarchical structure are sixteen root MeSH terms, such as “Anatomy”, “Diseases”, “Chemicals and Drugs”, and “Phenomena and Processes”. These broader terms maintain parent-child relationships in that each parent MeSH term branches into more specific “child” MeSH terms that fall under the umbrella of the broader “parent.” MeSH terms are not limited to any one branch, which means that MeSH terms can have multiple parents. For example, “Breast Neoplasms” has the parent “Neoplasms by Site” as well as “Breast Diseases” (Figure 4A). To reflect this hierarchical structure so that broader MeSH terms are mapped more promiscuously to articles than narrower MeSH terms, we normalized the frequency of MeSH terms (Figure 2B) by mapping the ancestors of a MeSH term back to articles of the descendants (Figure 4B). This ensures that broader, parent MeSH terms are mapped at higher or comparable frequencies than their narrower, more specific descendants. This same type of normalization can be seen in gene sets for hierarchical pathway datasets like that of Reactome [38, 39]. This normalization prevents skewing of results towards broader MeSH terms, which may have been inconsistently mapped to articles, and enables identification of more specific associations with child terms within the MeSH tree, which cuts down on the overall noise to identify the most relevant associations. This normalization is defined in equations and in Table 1.

2.4 Ranking Gene-MeSH co-occurrences

The naive GeneID-MeSH network only contains associations between a gene and a MeSH term based on the frequency with which those two entities occur together in an article. To extract meaningful co-occurrences of a GeneID and a MeSH term, we calculated a rank measure called normalized pointwise mutual information (NPMI), which is the normalized variant of pointwise mutual information (PMI) (Table 1). PMI is a rank measure commonly used in text mining for collocation extraction, i.e., identifying words that co-occur together more than random indicating a shared meaning like “hot tea” and “crystal clear”. Because PMI is a rank measure, there is no level of significance or accepted cutoff to use for co-occurring terms; however, the normalized variant, NPMI, calculates a continuous value between -1 and 1 where -1 is interpreted as no co-occurrence, 1 is interpreted as perfect co-occurrence, and 0 is interpreted as co-occurrence at random [44]. These interpretations can be made about GeneID- MeSH co-occurrences because GeneIDs were tagged to articles independent of MeSH terms. Also, NPMI is biased in that low frequency co-occurrences are ranked higher [44]. To reduce the potential for spurious or less-replicable co-occurrences to drive this bias, GeneID-MeSH associations with less than three PubMed articles were excluded from the network. This cutoff was chosen based on assumptions that can be made about positive reporting of results due to publication bias [49]. We assumed that for a GeneID-MeSH association with three or more PubMed articles that support the association, it was likely that at least two of the articles reported positive results for a relationship between the GeneID and MeSH term.

Table 1 summarizes the equations needed to calculate NPMI. The probability of a MeSH term and all the descendants of occurring () will increase the denominator of PMI resulting in an overall lower NPMI for broader terms since frequency is increased for a given MeSH term based on its descendants. This adjustment decreases the overall ranks of MeSH terms with many descendants whereas more specific MeSH terms ranked higher. The final network was filtered to include only GeneID-MeSH associations with $NPMI > 0$, which indicates that each association present exceeds the associations expected from random chance (Figure 2B).

2.5 MeSH Terms for breast carcinogenesis

MeSH terms that comprehensively capture the use case of breast carcinogenesis were needed to query EMCON and retrieve a ranked list of relevant genes. As described in Grashow et al. [32], experts selected seventeen MeSH terms that encompassed concepts from seminal papers on carcinogenesis [50–52] and breast carcinogenesis [9], including: Neovascularization, pathologic; Neovascularization, physiologic; Apoptosis; Cell cycle; Epigenomics; DNA damage; DNA repair; Growth hormone; Cell survival; Immune system; Inflammation; Breast; Breast Diseases; Oxidative stress; Cell proliferation; Gonadal steroid hormones; and Xenobiotics. These seventeen MeSH terms alone do not necessarily reflect the full scope of the concept they represent, so the full query also includes all descendants of these MeSH terms for a total of 214 MeSH terms to represent breast carcinogenesis. Clearly, some of these concepts may be related to cancer phenotypes more broadly, and some may be more specific for breast carcinogenesis.

2.6 Relevance of retrieved gene list

For the topic of breast carcinogenesis, a reference gene set of 287 genes was compiled through expert literature review (ELR) by Silent Spring Institute as described in Grashow et al. [32], including: (1) gene targets for quantitative nuclease protection assays in ToxCast Phase I; (2) genes responsive to nuclear receptors of interest (estrogen, progesterone, androgen, and aryl hydrocarbon receptors); (3) genes included in Qiagen microarray panels designed to probe pathways relevant to breast cancer (estrogen receptor signaling, breast cancer, DNA repair, DNA damage, growth factors, cellular stress response); (4) important genes in breast cancer based on key literature reports [4, 53–55]; (5) genes listed as related to breast cancer in curated databases (OMIM, CTD); and, (6) genes listed by partners at NCATS Chemical Genomics Center (NCGC) as important in cytotoxicity response (Figure 2C). Potential housekeeping genes were chosen from previous reports in MCF-7 cells [56–58]. This ELR gene set was used as a reference gene set to measure the relevance of the retrieved gene list from EMCON to the topic of breast carcinogenesis.

The EMCON search was conducted 214 times to generate one gene list for each MeSH term in the search query. The final gene list was obtained by averaging the NPMI rank per gene in the set of 214 iterations. Relevance to breast carcinogenesis of the final gene list from EMCON was measured by comparing the mean rank of the ELR gene set to the distribution of mean ranks of 1000 randomly generated gene sets of the same length as the ELR gene set of 287 genes. The retrieved gene list was considered relevant to breast carcinogenesis if the mean rank for the ELR gene set is higher than the distribution of mean ranks for randomly generated gene sets yielding an empirical p-value < 0.01 . We used an empirical p-value because the comparison dataset is simulated, i.e. it was not derived using reference gene sets from other disorders. We felt that the best comparison would be against random gene sets rather than make inferences about how similar or dissimilar disorders may be based on respective genes. Recall was calculated as the fraction of ELR genes retrieved in the final list produced by EMCON, where the ELR gene set was considered a standard to evaluate the gene list produced by EMCON. Precision scores were calculated based on expert assessment of relevance of the top five genes for each of the seventeen selected MeSH terms. This expert assessment involved manual review of the literature that resulted in the GeneID-MeSH association to classify the association as true positive or false positive.

2.7 Comparison to a similar tool

EMCON's performance was compared to Génie, another literature-based gene prioritization approach [59]. Génie uses a naive linear Bayesian classifier in conjunction with a Fisher's exact test to produce a list of genes ranked by false discovery rate (FDR). We compared our method with that of Génie by using a Spearman rank correlation of the ELR gene set from the EMCON search results with search results from Génie. Two gene sets were obtained from Génie: one using only the MeSH term "Breast Neoplasms" and another using all 214 MeSH terms used to query EMCON.

2.8 Computational and statistical analyses

All data were downloaded as flat files from their respective sources (Table 2). Python 3.6+ [60] was used to parse the files and import into MongoDB 3.4+ [61]. All methods were

implemented using MongoDB's aggregate pipeline or python packages pandas 0.20+ [62], numpy [63], and numba [64]. All code is available via iPython notebooks [65] at <https://github.com/USEPA/CompTox-HTTr-EMCON>.

3 Results

3.1 Entity MeSH co-occurrence network (EMCON)

Seven resources that manually tag PubMed articles with GeneIDs were identified (Table 2) and integrated into a single resource containing GeneID-PMID associations. Subsequently, MeSH terms were incorporated to generate a naive GeneID-MeSH network. Most of the genes in the naive GeneID-MeSH network are not human, but many produce proteins with high similarity to human protein orthologs, such that the information from non-human genes may be relevant to human-related research questions. To boost the number of articles mapped to human genes, UniRef50 clusters were used to identify human orthologous genes to increase the human relevant articles from around 500,000 to nearly 900,000. Next, the MeSH term frequency was normalized by mapping MeSH term ancestors back to articles to which their descendants were already mapped. Finally, GeneID-MeSH were ranked using NPMI to create a final network called Entity MeSH Co-occurrence Network (EMCON). EMCON is comprised of nearly 14 million associations, and, when filtered to require an article count > 2 , the associations were dramatically reduced with 3.56 million remaining associations. The GeneID-MeSH associations in EMCON have article counts ranging from three to 10,276. The NPMI scores range from -0.5 to 0.7 with a mean of 0.025 (Supplemental Figure 1) and 2.13 million GeneID- MeSH associations with $\text{NPMI} > 0$.

3.2 MeSH term frequency normalization

The MeSH term frequency normalization (represented as $p(m')$; See Methods) increased the promiscuity of MeSH tree terms based on descendants within the hierarchical trees, via mapping MeSH terms back to the articles of their descendants. The probability of a given MeSH term occurring in the naive GeneID-MeSH network ($p(m)$) increased with the number of descendants present in the network. This increase in promiscuity for broader MeSH terms corresponds to decreased NPMI for associated genes. Figure 5 demonstrates the probability of a MeSH term occurring before ($p(m)$) and after ($p(m')$) frequency normalization; the probability of the MeSH term co-occurring with the gene of interest ($p(g,m)$); and the associated NPMI scores for the GeneID-MeSH co-occurrences for two MeSH branches, "Cell Cycle" and "Skin Diseases".

First, $p(m)$ and $p(m')$ were compared for MeSH terms in the "Cell Cycle" and "Skin Diseases" branches (Figure 5A). $p(m)$ did not inversely decrease with the depth of the MeSH hierarchical tree for "Cell Cycle" or "Skin Diseases." This relationship implied that "Breast Neoplasms" was broader than "Skin Diseases" because the $p(m)$ for "Breast Neoplasms" ($p(m)=0.001$) was greater than the $p(m)$ for "Skin Diseases" ($p(m)=2.3e-5$). However, after MeSH term frequency normalization, $p(m')$ decreased as the depth increased for a given branch. For example, the $p(m')$ for "M Phase Cycle Checkpoints," a term representing increased depth within the "Cell Cycle" branch, was less than the $p(m')$ values associated with its ancestors. Figure 5A also shows that despite $p(m')$ decreasing as depth increased

within the “Cell Cycle” branch, the MeSH term “Cell Nucleus Division” was nearly absent from the network altogether with a $p(m')=2e-6$. Similarly, following frequency normalization, the probability of the MeSH term “Skin Diseases” occurring in the gene-curated literature was greater than the probability of observing “Triple Negative Breast Neoplasms.”

Increases in $p(m')$ correlated with decreases in NPMI, as illustrated in Figure 5B. In other words, for more promiscuous MeSH terms, the GeneID-MeSH term co-occurrence for that term was less likely to be specifically relevant for the specific topic of interest. When looking at the association between the MeSH branch, “Skin Diseases,” with epidermal growth factor receptor (EGFR), we see that the broader MeSH terms “Skin Diseases” and “Breast Diseases” had an NPMI < 0 (Figure 5B), indicating that these MeSH terms were less relevant specifically to breast carcinogenesis. The NPMI scores for the MeSH terms “Breast Neoplasms” and “Triple Negative Breast Neoplasms” co-occurring with EGFR remained at 0.1 and 0.21, respectively, because the $p(m')$ remained relatively similar to $p(m)$. The NPMI decreased for most MeSH terms with MAD2L1 where “Cell Division” and “Cell Nucleus Division” drop below zero, which are excluded from EMCON. However, the NPMI scores for “Mitosis” and “M Phase Cell Cycle Checkpoints” remained above 0, so these associations were preserved. Despite the decreased NPMI for “Cell Cycle” and MAD2L1, these associations remained above 0 and were also preserved.

By normalizing the MeSH term frequency, we reduced the noise introduced into the network to retrieve more specific and useful GeneID-MeSH co-occurrences. This network cleaning approach assured that broader terms would not be ranked higher than more specific terms. The net impact is that less-specific MeSH terms will have lower NPMIs; many articles relate to “Skin Diseases” or “Breast Neoplasms,” but these articles may have little association with “Triple Negative Breast Neoplasms.” MeSH terms more closely associated with the pathological finding of interest such as “Triple Negative Breast Neoplasms,” will have a greater NPMI due to closer association. For all gene ID-MeSH co-occurrences for a given branch, the NPMI will increase with depth; i.e., the lowest descendant MeSH term-gene co-occurrence will have the greatest NPMI. Thus, the most salient associations will be quantitatively identified.

3.3 Relevance of search results to breast cancer

Genes related to the topic of breast carcinogenesis were retrieved from EMCON using seventeen expert-selected MeSH terms that represent concepts from seminal papers on the topic of specifically breast carcinogenesis [9] and carcinogenesis in general [50–52]. These seventeen MeSH terms were expanded to include all descendants in the MeSH trees to ensure the full scope is represented within the selection. The final list of MeSH terms totals 214, which were used to query EMCON and retrieve a final list of 14,811 genes. The full breast cancer results matrix, including the NPMI rank for each gene, is provided as Supplemental File1.

Relevance of the EMCON-returned genes to breast cancer was evaluated by comparing the mean rank of the ELR gene set to the distribution of the mean ranks of randomly generated gene sets (Figure 6). The random gene sets were generated by randomly selecting 287 genes,

which is the length of the ELR gene set. The average rank of the ELR gene set was clearly distinguished from the random gene set distribution (empirical p-value $\ll 0.01$). Using the ELR gene set, recall from EMCON search results was 0.983. Precision was calculated by manually assessing the relevance of the top five genes with the corresponding MeSH term. The average precision across the seventeen selected MeSH terms was 0.87 (Table 3). We then looked at the top MeSH terms related to well-studied, breast cancer genes: *BRCA1*, *BRCA2*, *ESR1*, *ESR2*, and *PGR* (Table 4). The MeSH terms retrieved are all specific to breast cancer or molecules linked to breast cancer like “Progesterone” and “Estradiol”.

3.4 Comparison to Génie

We searched Génie with two different queries to obtain breast cancer-related genes to compare to EMCON results. The Spearman rank correlation for the results from the query with all 214 MeSH terms is 0.561 (Figure 7) with a recall for the ELR gene set of 0.718. When using only the MeSH term “Breast Neoplasms” to retrieve breast cancer-related genes, the Spearman rank correlation drops to 0.451 (Figure 7) and the recall for the ELR gene set also drops to 0.641.

4 Discussion

We have developed an accessible and scalable resource called EMCON that is comprised of ranked associations between genes and MeSH terms. This novel tool is a needed public health and toxicology resource that enables connection of an AO of concern with hypothetical MIE or KE information, thus improving development of putative AOPs and providing an empirical approach to hypothesis generation. EMCON was developed via integration of multiple data sources and subsequent computation of the rank of specific associations. In the example herein, a ranked list of genes putatively related to breast carcinogenesis was defined using EMCON for use in hypothesis testing. The performance of EMCON in this example was evaluated in three ways: (1) comparison of the mean rank of the ELR gene set compared to randomly generated gene sets from the EMCON search results using the expert-selected MeSH terms related to breast carcinogenesis; (2) evaluation of the recall and precision of the EMCON search results using the ELR-derived gene set as a standard; and, (3) comparison to results for the 214 breast carcinogenesis-related MeSH terms from an existing tool, Génie. These three evaluations demonstrated that EMCON performed well for the use case of defining genes linked to MeSH terms. Within the EMCON search results, the ELR gene set for breast carcinogenesis ranked, on average, higher than any randomly generated gene set based on NPML. Further, EMCON demonstrated excellent manually assessed precision (0.87) and recall (0.983) using the ELR gene set, and the EMCON results correlated with results from Génie, with some differences noted based on different methodological choices. Overall, the results presented herein suggest this is a valuable tool for hypothesis generation, providing critical support for the building of AOPs and AOP networks in addition to advancing research in biological and biomedical fields.

EMCON was constructed to better utilize existing information in systematic information extraction of information used in hypothesis generation. EMCON was built by first

integrating heterogeneous resources that map genes to articles containing information across a multitude of topics from PubMed. Then protein similarity clusters from UniRef50 were used to identify articles with similar, non-human genes to be mapped to the corresponding human gene. MeSH term frequency was normalized by mapping MeSH term ancestors back to articles of their descendants so that MeSH frequencies correspond to the depth of the MeSH tree. Lastly, GeneID-MeSH associations were ranked using NPMI. For construction of EMCON, we utilized several resources that manually curate PubMed articles with genes relevant to the content of the article. Each curation effort prioritizes articles based on specific areas of interest: pathways (Reactome), proteins (UniProt), chemical-gene/gene product interactions (CTD), etc. The total number of articles from all the resources totals to almost 1 million out of 27 million articles within PubMed. Without the development of systematic information extraction or tagging efforts like NER [30], researchers are forced to rely on manual approaches. The mappings from manual efforts may be higher quality than those derived from potential systematic approaches, but the throughput is low. Also, each resource is biased towards a specific topic, so specific topics of interest may not be well represented in EMCON.

An ongoing limitation of any data mining approach using manually curated information from PubMed is that curation efforts are not standardized, as demonstrated by curation of certain GWAS and gene expression profiling studies. These curated studies have varying number of genes mapped without an explanation of whether it was all genes included in the panels, only variants identified, or only those with differential expression. This lack of standardization or clarification introduces noise into the network. However, the article count cutoff used in this approach (see Methods) filtered out much of this noise, reducing the total set of associations by 75%. Noise in the set of GeneID-MeSH associations was also reduced through MeSH term frequency normalization, which is similar to approaches used in overrepresentation analysis like gene set enrichment analysis (GSEA) [1]. The Reactome gene set available for GSEA or similar pathway analysis methods is normalized in the same manner, i.e., genes from the child pathways are all annotated to the parent pathways as well [38]. A further limitation of only using manually curated information is applicability to certain use cases. In this work, we explored breast cancer, which has a lot of literature in the curated space, but other diseases or outcomes may not be associated with any curated data. In moving forward with this work, systematic approaches can be developed to extract relevant information from articles to fill in gaps in knowledge. Although a particular disease or outcome may have limited information, EMCON could be used if more broad MeSH terms could be connected to these topics.

The universe of possible human-relevant GeneID-MeSH associations was expanded by using UniProt Reference 50 (UniRef50) clusters to map non-human genes to corresponding human orthologs. Human genes are overrepresented in the curated PubMed literature accounting for nearly 50% of the articles with thousands of other species accounted for the remaining articles. However, genes from other research conducted in model species may be relevant to human pathogenesis, at least at the level of hypothesis generation. UniRef50 was used because the protein clusters included the cross-species orthologs whereas UniRef90 and UniRef100 are typically clusters of same-species protein isoforms. Homologene is a resource that also clusters cross-species gene orthologs together [66] and is used by similar

methods in literature-based gene prioritization [59]. However, Homologene has not been updated since the last release in 2014. UniRef50 is regularly updated and supports many other efforts in proteomics work, and thus presented a clear choice for use in EMCON. It is possible that by expanding the network to include human orthologs, we introduced noise by including genes that may not be relevant for human pathogenesis. This aspect could potentially be explored in further analyses, especially since these articles on the human orthologs can be easily identified and then removed if deemed irrelevant. This type of noise would not necessarily detract from the utility of EMCON for hypothesis generation.

Normalized pointwise mutual information or NPMI was chosen as the association measure because of the defined threshold of $NPMI > 0.0$ is interpreted as having dependent co-occurrence. Similar measures exist like Fisher's exact, log-likelihood ratio, and Pearson's chi-square [67], but despite their similar use in ranking, do not have defined thresholds to distinguish between independent and dependent co-occurrence. Similar co-occurrence measures to identify GeneID-MeSH associations are implemented in Gene2MeSH [68] and MeSHOPs [69]. However, Gene2MeSH does not normalize the MeSH term frequency, and both lack cross species similarity mappings and use Fisher's exact test to identify and rank MeSH terms associated with a gene. There is currently no consensus on which association measure works best because each measure can outperform the other depending the dataset [67, 70]. For the purposes of this paper, NPMI was chosen because a continuous rank measure could be more easily incorporated into other methods like the gene prioritization workflow implemented in Grashow et al. (2018). It is possible that the best use of an association measure with this dataset may be a combination of the previously listed measures. However, this work demonstrated the utility of NPMI for this problem where previous work has only focused on use of Fisher's exact [68, 69] or more complex machine learning methods [59].

The utility of EMCON was demonstrated within the scope of breast carcinogenesis. Breast carcinogenesis was chosen as a use case because of the large amount of information available on the topic due to major public health interest. Seventeen MeSH terms were selected by experts that, along with their descendants, represented important characteristics of breast carcinogenesis. A total of 214 MeSH terms were used to query EMCON and retrieve ranked lists of genes where the NPMI was averaged across all MeSH terms for a final ranked list of genes. The MeSH tree was not considered when MeSH terms were selected, so the depth of each MeSH term varies. Due to the inclusion of descendants with MeSH terms at varying depths, MeSH terms like "Immune System" are overrepresented with 69 descendants, while MeSH terms like "Epigenomics" are underrepresented with 0 descendants. This over- and underrepresentation introduces bias in the results for breast cancer. It is possible the MeSH term selection process can be improved with a systematic, data-driven approach rather than a manual approach. The MeSH term selection could also be improved by consideration of the parent-child relationships in the MeSH tree and the MeSH-MeSH association from common publications [71].

Comprehensive gene sets for complex disorders are typically comprised of variants derived from genome-wide association studies (GWAS). However, for this work, we wished to include a more heterogeneous set of genes that may be related to processes seen in

carcinogenesis. A breast cancer-specific gene set compiled through ELR was used to evaluate the relevance of the retrieved gene list to the topic of breast carcinogenesis. Using the ELR gene set as a standard, the final breast cancer gene list from EMCON had a recall of 0.983, and the ELR gene set ranked well above randomly generated gene sets of the same length (empirical $p < 0.01$) indicating that the higher-ranking genes from EMCON are likely relevant to breast carcinogenesis. This was further demonstrated by manually assessing precision of the top five genes for the seventeen selected MeSH terms (average precision=0.87). The MeSH terms that did not have a precision of 1 had either very few descendants (“Cell Survival” and “Epigenomics”) or had many descendants (“Immune System” and “Inflammation”; Supplemental Figure 2). MeSH terms with few descendants could represent newer topics with fewer relevant articles or could represent cases wherein few genes have been specifically annotated to the topic, e.g., “Epigenomics”. MeSH terms with many descendants could have genes promiscuously mapped to them because of the large number of varied topics within the descendants. In both cases, the genes that were not explicitly related to the corresponding MeSH term were not well annotated, demonstrated low article count relative to the other top- ranked genes, or may have resulted from promiscuous mapping of geneID to MeSH. This type of false positive is an artifact of using NPMI since rare co-occurrences (GeneID-MeSH associations with low article counts) are artificially ranked higher. The article count cutoff could be raised to a more conservative number to remove these types of associations and tune EMCON to the specific research application based on the level of specificity required. When evaluating MeSH terms related to well-known breast cancer-related genes (*BRCA1*, *BRCA2*, *ESR1*, *ESR2*, and *PGR*), the topics were all specific to breast cancer in that they related to breast tissue-specific tumors or molecules like estradiol and progesterone.

EMCON’s performance was compared to results from Génie [59]. The Spearman rank correlations indicate that EMCON has a strong positive correlation with a more complicated method. The recall values for the ELR gene set in both result sets from Génie were much lower than EMCON; Génie did not retrieve all genes identified as breast cancer-relevant through expert review. The differences in recall between the two tools may be due to some key differences in the function of Génie; unlike EMCON, Génie relies on gene2pubmed and GeneRIF [57] rather than all available sources of curated GeneID-PMID mappings and does not correct for MeSH term tagging frequency. EMCON is further distinguished from Génie as a standalone, easily searchable resource that is scalable to include updates from any of the included resources.

Other previous efforts in data mining to link genes to disease have included a variety of implementations, including GeneDistiller [72], Endeavor [73], and many more further outlined in Moreau and Tranchevent [74]. However, many of these resources have not been updated or maintained, are commercial products, or have limited accessibility for further customized integration. Further distinguishing EMCON from these resources is the use of MeSH term frequency normalization, orthologous genes, and the ease of scaling to include other relevant resources. Finally, EMCON is compatible with previous efforts at putative AOP development, but clearly different in its approach. Putative AOPs have been developed using frequent itemset mining [20] based on shared chemicals in ToxCast and CTD, in an effort to identify MIEs and KEs that may be relevant. In contrast, EMCON works in the

opposite direction, i.e. starting with the AO and its associated MeSH terms, with the goal of finding possible targets for an MIE or KE related to an AO of interest.

EMCON has been used as one of several data streams for a gene prioritization project to identify breast cancer gene sets for investigating the molecular mechanisms of mammary carcinogens [32]. Other potential applications of EMCON include defining reference gene sets for high throughput transcriptomics chemical screening efforts [26, 32]. It can be used alongside traditional pathway analysis, as it provides a means of linking differentially expressed genes to perturbations at higher levels of biological organization (tissue, organ, body, etc.). Also, due to the scalability of EMCON, other data can easily be incorporated from chemical and toxicity resources like PubChem [75], the Toxicity Reference Database [76], and ToxCast. These additional resources would provide associations between chemicals and biological entities (genes, pathways, *in vitro* and *in vivo* toxicity endpoints), further expanding the utility of EMCON for hypothesis generation, chemical hazard identification and prioritization, and putative AOP development. Ultimately, EMCON provides a scalable, comprehensive resource to strengthen empirical experimental design and systematic literature review via prioritization of hypotheses based on GeneID-MeSH associations. EMCON is a bioinformatic tool for public health and biomedical sciences that leverages the existing body of information on putative gene-outcome relationships to support research and improve health outcomes.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

References

1. Subramanian A, et al., Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*, 2005 102(43): p. 15545–50. [PubMed: 16199517]
2. Fridley BL, Jenkins GD, and Biernacka JM, Self-contained gene-set analysis of expression data: an evaluation of existing and novel methods. *PLoS One*, 2010 5(9).
3. Papatheodorou I, Oellrich A, and Smedley D, Linking gene expression to phenotypes via pathway information. *J Biomed Semantics*, 2015 6: p. 17. [PubMed: 25901272]
4. Nelson ER, et al., 27-Hydroxycholesterol links hypercholesterolemia and breast cancer pathophysiology. *Science*, 2013 342(6162): p. 1094–8. [PubMed: 24288332]
5. Hoover RN, et al., Adverse health outcomes in women exposed in utero to diethylstilbestrol. *N Engl J Med*, 2011 365(14): p. 1304–14. [PubMed: 21991952]
6. Hamajima N, et al., Alcohol, tobacco and breast cancer--collaborative reanalysis of individual data from 53 epidemiological studies, including 58,515 women with breast cancer and 95,067 women without the disease. *Br J Cancer*, 2002 87(11): p. 1234–45. [PubMed: 12439712]
7. Chlebowski RT, et al., Estrogen plus progestin and breast cancer incidence and mortality in the Women's Health Initiative Observational Study. *J Natl Cancer Inst*, 2013 105(8): p. 526–35. [PubMed: 23543779]
8. Shioda T, et al., Expressomal approach for comprehensive analysis and visualization of ligand sensitivities of xenoestrogen responsive genes. *Proc Natl Acad Sci U S A*, 2013 110(41): p. 16508–13. [PubMed: 24062438]
9. Schwarzman MR, et al., Screening for Chemical Contributions to Breast Cancer Risk: A Case Study for Chemical Safety Evaluation. *Environ Health Perspect*, 2015 123(12): p. 1255–64. [PubMed: 26032647]

10. Pirone JR, et al., Age-associated gene expression in normal breast tissue mirrors qualitative age-at-incidence patterns for breast cancer. *Cancer Epidemiol Biomarkers Prev*, 2012 21(10): p. 1735–44. [PubMed: 22859400]
11. Parada H Jr., et al., Race-associated biological differences among luminal A and basal-like breast cancers in the Carolina Breast Cancer Study. *Breast Cancer Res*, 2017 19(1): p. 131. [PubMed: 29228969]
12. Wu Y, et al., Risk factors contributing to type 2 diabetes and recent advances in the treatment and prevention. *Int J Med Sci*, 2014 11(11): p. 1185–200. [PubMed: 25249787]
13. Gwinn MR, et al., Chemical Risk Assessment: Traditional vs Public Health Perspectives. *Am J Public Health*, 2017 107(7): p. 1032–1039. [PubMed: 28520487]
14. Ankley GT, et al., Adverse outcome pathways: a conceptual framework to support ecotoxicology research and risk assessment. *Environ Toxicol Chem*, 2010 29(3): p. 730–41. [PubMed: 20821501]
15. Perkins EJ, et al., Adverse Outcome Pathways for Regulatory Applications: Examination of Four Case Studies With Different Degrees of Completeness and Scientific Confidence. *Toxicol Sci*, 2015 148(1): p. 14–25. [PubMed: 26500288]
16. IBCERCC. Breast Cancer and the Environment: Prioritizing Prevention. 2013 IBCERCC (Interagency Breast Cancer and Environmental Research Coordinating Committee).
17. Kavlock R, et al., Update on EPA's ToxCast program: providing high throughput decision support tools for chemical risk management. *Chem Res Toxicol*, 2012 25(7): p. 1287–302. [PubMed: 22519603]
18. Tice RR, et al., Improving the human hazard characterization of chemicals: a Tox21 update. *Environ Health Perspect*, 2013 121(7): p. 756–65. [PubMed: 23603828]
19. Mansouri K, et al., CERAPP: Collaborative Estrogen Receptor Activity Prediction Project. *Environ Health Perspect*, 2016 124(7): p. 1023–33. [PubMed: 26908244]
20. Oki NO and Edwards SW, An integrative data mining approach to identifying adverse outcome pathway signatures. *Toxicology*, 2016 350–352: p. 49–61.
21. Davis AP, et al., The Comparative Toxicogenomics Database: update 2017. *Nucleic Acids Res*, 2017 45(D1): p. D972–D978. [PubMed: 27651457]
22. King BL, et al., Ranking transitive chemical-disease inferences using local network topology in the comparative toxicogenomics database. *PLoS One*, 2012 7(11): p. e46524. [PubMed: 23144783]
23. Amberger JS, et al., OMIM.org: Online Mendelian Inheritance in Man (OMIM(R)), an online catalog of human genes and genetic disorders. *Nucleic Acids Res*, 2015 43(Database issue): p. D789–98. [PubMed: 25428349]
24. Subramanian A, et al., A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles. *Cell*, 2017 171(6): p. 1437–1452 e17. [PubMed: 29195078]
25. Merrick BA, Paules RS, and Tice RR, Intersection of toxicogenomics and high throughput screening in the Tox21 program: an NIEHS perspective. *Int J Biotechnol*, 2015 14(1): p. 7–27. [PubMed: 27122658]
26. Mav D, et al., A hybrid gene selection approach to create the S1500+ targeted gene sets for use in high-throughput transcriptomics. *PLoS One*, 2018 13(2): p. e0191105. [PubMed: 29462216]
27. Coordinators NR, Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*, 2017.
28. NLM, Chapter 11 Relationships in Medical Subject Headings. 2016.
29. Hirschman L, et al., Overview of BioCreative IV: critical assessment of information extraction for biology. *BMC Bioinformatics*, 2005 6 Suppl 1: p. S1. [PubMed: 15960821]
30. Smith L, et al., Overview of BioCreative II gene mention recognition. *Genome Biol*, 2008 9 Suppl 2: p. S2.
31. The UniProt C., UniProt: the universal protein knowledgebase. *Nucleic Acids Res*, 2017 45(D1): p. D158–D169. [PubMed: 27899622]
32. Grashow RG, et al., BCScreen: A gene panel to test for breast carcinogenesis in chemical safety screening. *Computational Toxicology*, 2018 5(Supplement C): p. 16–24. [PubMed: 31218268]
33. Gene. [cited 2018 Jan 5]; Available from: <ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/gene2pubmed.gz>.

34. Brown GR, et al., Gene: a gene-centered information resource at NCBI. *Nucleic Acids Res*, 2015 43(Database issue): p. D36–42. [PubMed: 25355515]
35. GeneRIF. Gene Reference into Function. Available from: <https://www.ncbi.nlm.nih.gov/gene/about-generif>.
36. CTD. [cited 2017 11 30]; Available from: <http://ctdbase.org/downloads/>.
37. UniProt. [cited 2017 30 11]; Available from: <http://www.uniprot.org/downloads>.
38. Reactome. [cited 2018 1 05]; Available from: <https://reactome.org/>.
39. Fabregat A, et al., The Reactome Pathway Knowledgebase. *Nucleic Acids Res*, 2017.
40. RGD. [cited 2018 11 30]; Available from: ftp://ftp.rgd.mcg.edu/pub/data_release/.
41. Shimoyama M, et al., The Rat Genome Database 2015: genomic, phenotypic and environmental variations and disease. *Nucleic Acids Res*, 2015 43(Database issue): p. D743–50. [PubMed: 25355511]
42. Eppig JT, et al., Mouse Genome Informatics (MGI): Resources for Mining Mouse Genetic, Genomic, and Biological Data in Support of Primary and Translational Research. *Methods Mol Biol*, 2017 1488: p. 47–73. [PubMed: 27933520]
43. MGI. [cited 2017 11 30]; Available from: <http://www.informatics.jax.org/downloads/reports/index.html>.
44. Bouma G, Normalized (pointwise) mutual information in collocation extraction. *Proceedings of the International Conference of the German Society for Computational Linguistics and Language Technology*, 2009.
45. GeneRIF: Gene Reference into Function. [cited 2017 11 30]; Available from: <https://www.ncbi.nlm.nih.gov/gene/about-generif>.
46. Pfeffer U, et al., Estrogen receptor variant messenger RNA lacking exon 4 in estrogen- responsive human breast cancer cell lines. *Cancer Res*, 1993 53(4): p. 741–3. [PubMed: 7916651]
47. Kang HY, et al., Cloning and characterization of human prostate coactivator ARA54, a novel protein that associates with the androgen receptor. *J Biol Chem*, 1999 274(13): p. 8570–6. [PubMed: 10085091]
48. Suzek BE, et al., UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics*, 2015 31(6): p. 926–32. [PubMed: 25398609]
49. Song F, et al., Extent of publication bias in different categories of research cohorts: a meta-analysis of empirical studies. *BMC Med Res Methodol*, 2009 9: p. 79. [PubMed: 19941636]
50. Goodson WH 3rd, et al., Assessing the carcinogenic potential of low-dose exposures to chemical mixtures in the environment: the challenge ahead. *Carcinogenesis*, 2015 36 Suppl 1: p. S254–96. [PubMed: 26106142]
51. Hanahan D and Weinberg RA, Hallmarks of cancer: the next generation. *Cell*, 2011 144(5): p. 646–74. [PubMed: 21376230]
52. Smith MT, et al., Key Characteristics of Carcinogens as a Basis for Organizing Data on Mechanisms of Carcinogenesis. *Environ Health Perspect*, 2016 124(6): p. 713–21. [PubMed: 26600562]
53. Bastien RR, et al., PAM50 breast cancer subtyping by RT-qPCR and concordance with standard clinical molecular markers. *BMC Med Genomics*, 2012 5: p. 44. [PubMed: 23035882]
54. Cancer Genome Atlas N, Comprehensive molecular portraits of human breast tumours. *Nature*, 2012 490(7418): p. 61–70. [PubMed: 23000897]
55. Latimer JJ, et al., Nucleotide excision repair deficiency is intrinsic in sporadic stage I breast cancer. *Proc Natl Acad Sci U S A*, 2010 107(50): p. 21725–30. [PubMed: 21118987]
56. Chen L, et al., The enhancement of cancer stem cell properties of MCF-7 cells in 3D collagen scaffolds for modeling of cancer and anti-cancer drugs. *Biomaterials*, 2012 33(5): p. 1437–44. [PubMed: 22078807]
57. Chua SL, et al., UBC and YWHAZ as suitable reference genes for accurate normalisation of gene expression using MCF7, HCT116 and HepG2 cell lines. *Cytotechnology*, 2011 63(6): p. 645–54. [PubMed: 21850463]
58. Jacob MD, et al., Environmental cues induce a long noncoding RNA-dependent remodeling of the nucleolus. *Mol Biol Cell*, 2013 24(18): p. 2943–53. [PubMed: 23904269]

59. Fontaine JF, et al., Genie: literature-based gene prioritization at multi genomic scale. *Nucleic Acids Res*, 2011 39(Web Server issue): p. W455–61. [PubMed: 21609954]
60. Guido, Python tutorial, Technical Report CS-R9526. 1995.
61. MongoDB. Available from: <https://www.mongodb.com/>.
62. pandas: Python Data Analysis Library. 2012.
63. Van Der Walt S, Colbert C, and Varoquaux G, The NumPy array: a structure for efficient numerical computation. 2011.
64. Lam S, Pitrou A, and Seibert S. Numba: A LLVM-based Python JIT Compiler. in *Proceedings of the Second Workshop on the LLVM Compiler Infrastructure in HPC*. 2015 Austin, Texas: ACM.
65. Pérez F and Granger B, IPython: A System for Interactive Scientific Computing. *Computing in Science & Engineering*, 2007 9(3): p. 21–29.
66. Homologene. [cited 2017 11 30]; Available from: <https://www.ncbi.nlm.nih.gov/homologene>.
67. Pecina P, Lexical association measures and collocation extraction. *Language Resources and Evaluation*, 2010 44(1): p. 137–158.
68. Ade A, Wright Z, and States D. Gene2MeSH. 2007; Available from: <http://gene2mesh.ncibi.org>.
69. Cheung WA, Ouellette BF, and Wasserman WW, Quantitative biomedical annotation using medical subject heading over-representation profiles (MeSHOPs). *BMC Bioinformatics*, 2012 13(1): p. 249. [PubMed: 23017167]
70. Pearce D, A Comparative Evaluation of Collocation Extraction Techniques, in *International Conference on Language Resources and Evaluation*. 2002 p. 1530–1536.
71. Kastrin A, Rindflesch TC, and Hristovski D, Large-scale structure of a network of co- occurring MeSH terms: statistical analysis of macroscopic properties. *PLoS One*, 2014 9(7): p. e102188. [PubMed: 25006672]
72. Seelow D, Schwarz JM, and Schuelke M, GeneDistiller--distilling candidate genes from linkage intervals. *PLoS One*, 2008 3(12): p. e3874. [PubMed: 19057649]
73. Tranchevent LC, et al., Candidate gene prioritization with Endeavour. *Nucleic Acids Res*, 2016 44(W1): p. W117–21. [PubMed: 27131783]
74. Moreau Y and Tranchevent LC, Computational tools for prioritizing candidate genes: boosting disease gene discovery. *Nat Rev Genet*, 2012 13(8): p. 523–36. [PubMed: 22751426]
75. Kim S, et al., PubChem Substance and Compound databases. *Nucleic Acids Res*, 2016 44(D1): p. D1202–13. [PubMed: 26400175]
76. Martin MT, et al., Profiling the reproductive toxicity of chemicals from multigeneration studies in the toxicity reference database. *Toxicol Sci*, 2009 110(1): p. 181–90. [PubMed: 19363143]

Highlights:

- Biomedical text resources were integrated to identify gene-disease associations
- Normalized pointwise mutual information was used to identify and rank genes linked to key carcinogenic characteristics
- A relevant breast cancer gene set ranked higher than random gene sets
- Methods scale to include other biological and chemical concepts

Figure 1: MeSH tree branches for "Carcinoma, Ductal, Breast".

Shown are two branches for the MeSH term "Carcinoma, Ductal, Breast". These branches have two root MeSH terms: "Neoplasms" and "Skin and Connective Tissue Diseases". Preceding MeSH terms (i.e. traversal towards a root MeSH term) are ancestors, where descendants are the MeSH following (i.e. traversal away from a root MeSH term). The depth of a MeSH term corresponds to the number of ancestors it has, so depth increases with traversal away from the root MeSH term.

Figure 2: Workflow for building Entity MeSH Co-occurrence Network (EMCON).

EMCON is created by (A) integration of biomedical resources, specifically manually annotated datasets of GeneID-PMID mappings. These data are combined with PMID-MeSH mappings to create a naive GeneID-MeSH network. (B) Next, the naive GeneID-MeSH network is expanded by mapping orthologous genes followed by MeSH term frequency normalization. The GeneID- MeSH associations are then ranked to generate the final EMCON resource. (C) EMCON can be queried with specific use cases where experts identify MeSH terms important to a topic of interest. Those selected MeSH terms are expanded to include descendants, which is the full set of MeSH terms used to query EMCON. The final output is a ranked list of genes ranked according to overrepresentation of the topic of interest.

Figure 3: Example of the naive GeneID-MeSH network.

The naive GeneID-MeSH network consists of GeneIDs that have been manually tagged to articles within PubMed, which are connected to MeSH terms.

Figure 4: Example of MeSH term frequency normalization.

(A) Shown are the same branches from Figure 1. The MeSH term "Breast Neoplasms" has a total of five ancestors with two root MeSH Terms: "Neoplasms" and "Skin and Connective Tissue Diseases." (B) All the ancestors for a given a MeSH term are subsequently tagged to each article of a specific MeSH term. The original mapped associations are indicated by solid arrows and inferred associations as part of our MeSH term normalization are indicated by dashed arrows.

*MeSH terms only used for structuring the MeSH tree and not used for tagging articles.

Figure 5:
Two branches from the MeSH hierarchical tree were used to demonstrate how the annotation bias correction alters the probability of a MeSH term occurring ($p(m)$) along with the resulting NPMI with a relevant gene. These values correspond with the depth of a given MeSH term in the hierarchical tree.

Figure 6: Comparison of mean rank of ELR, breast cancer-specific gene set to random gene sets within EMCON search results.

The ELR gene set is, on average, ranked higher than any of the mean ranks for randomly generated gene sets of the same length. Shown are 300 representative random gene sets from a total of 1000. The mean rank across all the random gene sets is 7405.

Figure 7:

The rank comparison of the ELR gene set from EMCON and Génie. We obtained the two Génie gene sets by searching with 214 breast cancer-related MeSH terms and with only "Breast Neoplasms". The correlation of the rank comparisons was similar across the two queries.

Table 1:

Equations for ranking GeneID-MeSH co-occurrences.

Equation	Description
$C = \{c_1, \dots, c_n\}$	a set of co-occurrences, where c is a GeneID-MeSH term co-occurrence that is unique by PMID and n is the total number of co-occurrences
$G(g)$	the number of co-occurrences of C that contain gene, g
$M(m)$	the number of co-occurrences of C that contain MeSH term, m
$M(m')$	the number of co-occurrences of C that contain m and all the descendants of m
$T(g, m)$	the subset of C that contains co-occurrences with both g and m
$p(g) = \frac{ G(g) }{n}$	the probability of g occurring
$p(m) = \frac{ M(m) }{n}$	the probability of m occurring based on frequencies before MeSH term frequency normalization
$p(m') = \frac{ M(m') }{n}$	the probability of m and all the descendants of m occurring based on frequencies after MeSH term frequency normalization
$p(g, m) = \frac{ T(g, m) }{n}$	the probability of g and m co-occurring
$pmi(g, m) = \log\left(\frac{p(g, m)}{p(g)p(m')}\right)$	pointwise mutual information for a given g and m
$npmi(g, m) = \frac{pmi(g, m)}{-\log(p(g, m))}$	normalized pointwise mutual information for a given g and m

Table 2:
Manually curated resources used to construct EMCON.

A total of seven resources that manually tag GeneID's to articles within PubMed were integrated to serve as the initial dataset for building EMCON. Over 1.2 million articles make up the naive GeneID-MeSH network with over 7 million genes for over 14K species.

Gene and gene product databases	Number of articles	Number of GeneIDs in articles	Number of species across GeneIDs
gene2pubmed [33]	1,062,713	5,565,651	12,782
Gene Reference into Function (generif) [45]	705,441	90,329	1,913
Comparative Toxicogenomics Database (CTD) [36]	58,180	43,298	76
Universal Protein Resource (UniProt/Swiss-Prot) [37]	950,989	5,156,248	12,555
Reactome [38]	15,650	11,110	9
Rat Genome Database (RGD) [40]	834,585	87,874	7
Mouse Genome Informatics (MGI) [43]	181,519	42,020	1
Total Unique	1,238,879	7,074,406	14,126

Table 3:
Manual Precision for 17 selected MeSH terms.

Relevance of the five top ranked genes for each of the seventeen selected MeSH terms relevant to breast carcinogenesis was evaluated by performing a literature search of through Entrez Gene. Gene symbols in red were not explicitly related to the corresponding MeSH term.

MeSH name	Top five genes (gene symbol)	Precision
Neovascularization, Pathologic	<i>VEGFA, KDR, ANGPT2, ANGPT4, VASH1</i>	1
Neovascularization, Physiologic	<i>KDR, FLT1, TEK, ANGPT1, EPHB4</i>	1
Apoptosis	<i>CASP3, BAX, CASP9, BCL2, CASP8</i>	1
Cell Cycle	<i>CDK2, CDK1, CCNE1, CCNA2, CDKN1B</i>	1
Epigenomics	<i>PAP12, DNMT3A, TET3, GREB1, KAT8</i>	0.6
DNA Damage	<i>ATR, CHEK1, ATM, MDC1, DDB2</i>	1
DNA Repair	<i>RAD51, XRCC1, XPC, ERCC2, XPA</i>	1
Growth Hormone	<i>CSHL1, GH1, GH2, GHR, CSH1</i>	1
Cell Survival	<i>ARIH2OS, CASP3, BAD, BCL2, BCL2L1</i>	0.8
Immune System	<i>LAT2, CLEC4E, ARL4C, CLEC6A, NAV1</i>	0.6
Inflammation	<i>NLRP3, CRP, PYDC1, SPATA31E1, NLRP13</i>	0.8
Breast	<i>SCGB3A1, WISP3, PTK6, SCGB2A1, SCGB2A2</i>	1
Breast Diseases	<i>TBX3, IGFBP3, TP63, TP73, IGF1</i>	1
Oxidative Stress	<i>CAT, GSR, NFE2L2, SOD2, GPX1</i>	1
Cell Proliferation	<i>CCND1, FOXM1, CDKN1B, YAP1, CDKN1A</i>	1
Gonadal Steroid Hormones	<i>SEMG2, ACRV1, HSD17B1, SEMG1, HSD17B3</i>	1
Xenobiotics	<i>NR1I3, NR1I2, ACSM2A, ACSM2B, NAT1</i>	1
		0.870588

Table 4:
Top MeSH terms for genes *BRCA1*, *BRCA2*, *ESR1*, *ESR2*, and *PGR* from EMCON.

Five breast cancer-related genes were used to search EMCON. Shown are the top-ranking co- occurring MeSH terms.

GeneID	Gene Symbol	MeSH Term
672	<i>BRCA1</i>	Breast Neoplasms
		Triple Negative Breast Neoplasms
		Breast Neoplasms, Male
		Hereditary Breast and Ovarian Cancer Syndrome
		Carcinoma, Ductal, Breast
675	<i>BRCA2</i>	Breast Neoplasms, Male
		Hereditary Breast and Ovarian Cancer Syndrome
		Breast Neoplasms
		Triple Negative Breast Neoplasms
2099	<i>ESR1</i>	Estradiol
		Fibrocystic Breast Disease
		Estrogens, Conjugated (USP)
		Breast Neoplasms
2100	<i>ESR2</i>	Estradiol
		Estrogens, Conjugated (USP)
5241	<i>PGR</i>	Progesterone