Published in final edited form as:
Comput Toxicol. 2018 February ; 5: 16–24. doi:10.1016/j.comtox.2017.11.003.

BCScreen: A gene panel to test for breast carcinogenesis in chemical safety screening

Rachel G. Grashow[a], **Vanessa Y. De La Rosa**[a,b], **Sean M. Watford**[c], **Janet M. Ackerman**[a], and **Ruthann A. Rudel**[a,*]

[a]Silent Spring Institute, 320 Nevada Street, Newton, MA 02460, United States

[b]Social Science Environmental Health Research Institute, Northeastern University, Boston, MA, United States

[c]Department of Environmental Sciences and Engineering, Gillings School of Global Public Health, UNC-Chapel Hill, Chapel Hill, NC, United States

Abstract

Targeted gene lists have been used in clinical settings to specify breast tumor type, and to predict breast cancer prognosis and response to treatment. Separately, panels have been curated to predict systemic toxicity and xenoestrogen activity as a part of chemical screening strategies. However, currently available panels do not specifically target biological processes relevant to breast development and carcinogenesis. We have developed a gene panel called the Breast Carcinogen Screen (BCScreen) as a tool to identify potential breast carcinogens and characterize mechanisms of toxicity. First, we used four seminal reviews to identify 14 key characteristics of breast carcinogenesis, such as apoptosis, immunomodulation, and genotoxicity. Then, using a hybrid data and knowledge-driven framework, we systematically combined information from whole transcriptome data from genomic databases, biomedical literature, the CTD chemical-gene interaction database, and primary literature review to generate a panel of 500 genes relevant to breast carcinogenesis. We used normalized pointwise mutual information (NPMI) to rank genes that frequently co-occurred with key characteristics in biomedical literature. We found that many genes identified for BCScreen were not included in prognostic breast cancer or systemic toxicity panels. For example, more than half of BCScreen genes were not included in the Tox21 S1500+ general toxicity gene list. Of the 230 that did overlap between the two panels, representation varied across characteristics of carcinogenesis ranging from 21% for genes associated with epigenetics to 82% for genes associated with xenobiotic metabolism. Enrichment analysis of BCScreen identified pathways and processes including response to steroid hormones, cancer, cell cycle, apoptosis, DNA damage and breast cancer. The biologically-based systematic approach to gene prioritization demonstrated here provides a flexible framework for creating disease-focused gene panels to support discovery related to etiology. With validation, BCScreen may also be useful for toxicological screening relevant to breast carcinogenesis.

[*]Corresponding author. Rudel@silentspring.org (R.A. Rudel).

Conflict of Interest
The authors declare no conflicts of interest.

Appendix A. Supplementary data
Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.comtox.2017.11.003.

Keywords

Breast cancer; Chemical carcinogenesis; Xenoestrogen; Endocrine disrupting chemical; Normalized pointwise mutual information

Introduction

Breast cancer represents a significant public health concern, with approximately 250,000 new diagnoses in US women each year [2]. While heritable genetic mutations like *BRCA1* and *BRCA2* have been shown to greatly increase risk in a subset of the population [28,27,26], known non-heritable risk factors for breast cancer include exposure to pharmaceutical hormones, medical radiation, age of first birth and other aspects of reproductive history, post-menopausal body mass index (BMI), reduced physical activity, alcohol consumption, and smoking [34,1,46,25,23,38]. To date, most breast cancer research has focused on treatment, heritable gene mutations, and the behavioral factors mentioned above. However, hormone-relevant risk factors and other medical and scientific evidence suggest additional influences on non-heritable breast cancer risk, including environmental chemical exposure [14,18,38]. At least three environmental chemical classes that are likely to increase breast cancer risk have been identified: 1) chemicals that cause mammary gland tumors via DNA damage pathways, 2) endocrine disrupting chemicals (EDCs) that alter mammary tumor growth, 3) toxicants that alter susceptibility by disrupting mammary gland development [7,41,39,38]. Investigating environmental chemical risk and associated mechanisms of carcinogenesis could inform prevention efforts, shape public health policy, and also illuminate avenues for new treatments.

Currently, the Mammary Carcinogens Review Database names over 200 chemicals considered to be mammary carcinogens (MCs) based on increased mammary gland tumors in animal studies [40,39]. However, many commercial and industrial chemicals have not been evaluated for breast cancer risk at any level. With thousands of untested chemicals in use and many more being introduced to the market each year, current toxicological approaches are not sufficient to identify chemicals that may increase breast cancer risk.

A paradigm proposed by the Interagency Breast Cancer and Environmental Research Coordinating Committee (IBCERCC) and others suggests working " … backward from a disease to identify the changes caused by chemicals that could serve as early indicators of toxicity" [9,24,43]. Such changes may serve as a link between cellular events relevant to breast cancer that are also responsive to environmental chemical exposures. For example, *in vivo* and *in vitro* studies have shown that environmental exposure to carcinogens and endocrine disruptors may exert influence via changes in gene expression [20,53,49,47]. These and other gene expression changes could be used to build a centralized list of environmentally susceptible genes that are also important in breast cancer. Such a gene list would serve as a critical tool in the evaluation of chemicals for carcinogenicity, and advance our mechanistic understanding of mammary carcinogens and mammary gland developmental disruptors.

Gene prioritization is broadly defined as the process by which the most promising genes or proteins are selected or targeted from a larger pool using systematic methods [32]. Some involve the use of a "seed" or training genes that are already associated with the endpoint or process of interest [52], while others are developed through the manual review of public biomedical and scientific databases. Targeted gene panels have previously been used in a variety of prediction contexts: to identify breast tumor subtype [10], therapeutic response [11], and likelihood of tumor recurrence [37]. Separately, a number of toxicology initiatives have sought to create sentinel or representative gene sets that can serve as markers or predictors of systemic toxicity including Tox21′s S1500+ [31], and the LINCS L1000 list [36,16]. Finally, gene panels have also been created to predict estrogen activity as a tool to identify xenoestrogens [42]. However, none of the currently available gene panels specifically target diverse biological processes relevant to breast development and carcinogenesis. This gap may be due to the difficulty in integrating multiple sources of gene data, as well as the heterogeneity of data quality and database curation.

To this end, we have developed a framework to prioritize and curate a panel of 500 genes to serve as a biomarker of mammary toxicity and breast carcinogenesis. Known as the Breast Carcinogen Screen (BCScreen), this approach represents a departure from conventional breast cancer gene platforms which focus on biomarkers for diagnosis and response to therapeutics. Instead, BCScreen is intended as a tool to identify potential breast carcinogens and the key molecular initiating events and pathways that may increase breast cancer risk in the context of chemical exposure. It can be applied in experimental studies in animals or in vitro systems, including high-throughput chemical screening. BCScreen synthesizes information from multiple sources including seminal papers on carcinogenesis, primary biomedical literature, whole transcriptome data from the publicly available GEO database, chemical-gene interactions from the Comparative Toxicogenomics Database (CTD) and expert literature review (ELR). In addition, this study introduces a novel application of normalized pointwise mutual information (NPMI), a data-mining technique that assigns a co-occurrence score between genes in PubMed and 14 key characteristics of breast carcinogenesis [55]. This framework combines data and knowledge driven approaches in that it relies on expert judgment for input selection and weighting, and subsequently applies a systematic approach to select genes based on those criteria and weights. This hybrid methodology allows us to integrate heterogeneous data, while maintaining flexibility to accommodate alternate input streams, model systems and weighting schema relevant to breast carcinogenesis or other disease etiologies.

Methods

In order to select genes for BCScreen, we first integrated four informative data streams including: 1) genes annotated in the scientific literature to biological processes established as key characteristics of carcinogenesis using normalized pointwise mutual information (NPMI); 2) gene expression data from endocrine disruptor exposures in vitro; 3) genes associated with rodent mammary carcinogens or chemicals that alter rodent mammary gland development; and 4) genes identified as involved in breast carcinogenesis through traditional expert literature review (ELR). Genes from these data sources were united into a candidate list of 18,482 genes, intended to capture all genes that may be relevant for carcinogenesis and mammary cancers. Fig. 1 shows the sources and selection criteria for candidate genes. Criteria and methods for ranking genes and identifying 500 high priority genes are described below.

Key characteristics of breast carcinogenesis

To capture key biological characteristics important in breast carcinogenesis, we identified seminal papers on carcinogenesis [21,19,48] and breast cancer [43]. Reviewing these articles, we compiled 14 common characteristics or biological processes important in breast carcinogenesis (Table 1). Many characteristics were identified by multiple sources. For example, evading apoptosis, immune modulation, inflammation, genotoxicity and immortalization were identified in all of the seminal papers. The mammary gland was only specifically mentioned in Schwarzman et al. [43]. Each characteristic was then assigned a Medical Subject Heading (MeSH), which represents a controlled vocabulary of over 27,000 keywords structured in a hierarchical tree to categorize concepts covered in an article. MeSH terms are both manually and automatically tagged in articles in PubMed, a bibliographic database largely comprised of biomedical literature maintained by the US National Library of Medicine (NLM).

Normalized pointwise mutual information (NPMI) scoring to identify genes associated with characteristics of carcinogens in biomedical literature

To associate and rank genes for each characteristic, we extracted relevant gene-MeSH associations using normalized pointwise mutual information (NPMI; [8]). NPMI is a text-mining algorithm that identifies and ranks word pairs that co-occur more frequently like “hot tea” or “crystal clear”. Specifically, NPMI assigns a rank measure between −1 and 1, such that −1 means no co-occurrence, 1 means perfect co-occurrence, and 0 means co-occurrence at random. A gene-MeSH term association network was created by integrating gene-curated articles from multiple resources including PubMed, the Comparative Toxicogenomics Database (CTD), the Rat Genome Database (RGD), Mouse Genome Informatics (MGI), and Universal Protein Resource (UniProt). This network paired every gene mentioned in these sources with each of the MeSH terms selected to represent the 14 key characteristics (Table 1).

We identified genes that are overrepresented for each of the characteristics by selecting those genes with an NPMI greater than 0. This meant that each gene has evidence of co-occurrence with one or more of the breast cancer-related MeSH terms over random chance [55]. In total, 14,545 genes had a greater than zero ranking with at least one characteristic. Genes lacking an association with a characteristic were assigned an NPMI of 0.0 for that characteristic. For characteristics with multiple MeSH terms, NPMIs were averaged across terms. UniProt Reference Clusters (UniRef50) was used to identify human homologs of non-human genes identified in the gene-MeSH network with at least 50% sequence identity and at least 80% overlap [3,50]. Human homologs increased the total number of relevant articles from under 500K to over 700K curated articles.

Gene expression changes in MCF-7 cells treated with endocrine disruptors

To incorporate genes with evidence of expression changes in response to endocrine disruptor exposure, we used NCBI's Gene Expression Omnibus (GEO; [17,4]). GEO is a publicly available data repository where researchers deposit and retrieve microarray and other functional genomics data. We combined 13 transcriptome-wide datasets (Affymetrix Human Genome U133 Plus 2.0 Array only) from NCBI's GEO database describing MCF-7 cells treated with endocrine disrupting chemicals (GSE5200: [44]; GSE7765: [22]; GSE50705: [45]). Chemicals include Bisphenol A (BPA), daidzein, diethylstilbestrol (DES), 17β-estradiol (E2), ethinyl estradiol (EE2), genistein, p-nonyl phenol (PNP), tris 4-hydroxyphenyl-4-propyl-1-pyrazole (PPT), and dioxin. When chemical dose-response data were available, we selected gene expression results at IC10-IC20 concentrations range for cell proliferation. This approach was used to include genes that are responsive at low dose concentrations. These efforts produced a list of 13,256 genes that showed significantly altered expression using the Benjamini and Hochberg false discovery rate ($p < 0.05$; [6]).

Genes associated with mammary carcinogens and mammary gland developmental disruptors

Previously, we identified 243 chemicals that are either mammary carcinogens, mammary gland development disruptors or both (Supplemental Table 1; [40,41]). We used this list to probe the chemical-gene interaction data maintained by the Comparative Toxicogenomics Database (CTD; [15]). CTD uses a manually curated literature pool of approximately 50,000 articles with information related to chemical-gene/protein interactions in vertebrates and invertebrates. The mammary carcinogen list contained 237 chemicals that matched to any chemical-gene interaction in CTD. We selected those chemical-gene interactions that were identified in mammals or zebrafish models and had greater than three publications supporting the interaction. Human homologs for non-human genes were identified using UniRef50 [3,50]. The final list of genes from CTD included 15,078 unique entries that were linked to 99 chemicals from the mammary carcinogen list with sufficient chemical-gene interactions.

Expert literature review (ELR)

We used a traditional subject matter expert literature review to manually curate a list of 289 genes important in normal breast biology and breast cancer etiology and progression. The ELR was considered to contain high value observations that may not yet have been transmitted to literature or other toxicological databases. This expert-based list includes genes important in estrogen, progesterone, and prolactin activity, breast-specific xenobiotic metabolism [56,33], genotoxicity, rodent mammary gland biology [60], breast tumor biology and biomarkers of breast cancer prognosis identified from several sources including Quantitative Nuclease Protection Assays (qNPA) in ToxCast [5], in vitro cancer hallmarks from Toxcast [58], Qiagen breast cancer microarray panels, BC-related literature [30,10,54], and curated databases (OMIM, CTD, tumorgene.org).

Summary score and BCScreen gene panel selection

We combined the four candidate gene lists, calculated a summary score and ranked each gene as described below and depicted in Fig. 2. To calculate the summary score for each gene, the NPMI score for each gene in each of the 14 categories – which ranged from 0 to 1 - was incrementally increased by 0.5 if it appeared on the GEO dataset, by 0.5 if it appeared on the ELR list and by 0.05 if it resulted from the CTD search. The GEO and ELR datasets were equally weighted to ensure that the dataset was enriched for MCF-7 genes that have shown expression changes in response to exogenous compounds and for genes identified through expert knowledge sources in breast carcinogenesis. Genes from the CTD dataset were not weighted as heavily for a number of reasons. First, the gene-chemical associations were based on a relatively small number of manually curated articles, so relevant associations may be missing (see Discussion). In addition, the CTD data incorporated into the gene prioritization algorithm did not specify the nature of the gene-chemical interaction. Finally, some of the gene-chemical interactions may be indirect, i.e. mediated by other chemicals or genes. Each summary score was assigned a random number between 1E-10 and 5E-11 to break any ties.

To compile a panel of 500 genes for BCScreen, each gene was assigned to the cancer characteristic where it held the highest summary score, and then the top 33 genes were selected for each characteristic (71 for mammary). Once a gene was assigned to the characteristic where it had the highest score, that gene could not be selected for any other characteristic. This process was repeated until each characteristic contained a unique set of the highest scoring genes, resulting in a final panel of 500 genes (see Supplementary Table 2).

Enrichment analysis

To evaluate whether BCScreen captured biological processes and terms relevant to the 14 characteristics, we conducted enrichment analysis using the web-based Enrichr software (http://amp.pharm.mssm.edu/Enrichr/; [12,29]). Enrichr uses a compendium of over 70 publicly available libraries containing ontologies, gene-disease and gene-metabolism information to rank pathways and terms in user-up-loaded gene lists. Enrichr presents significance using multiple measures including the Fisher exact test as well as an in-house rank-based statistical measure. We selected the following nine pathway or term libraries from the Enrichr portal: Disease_perturbations_from_GEO_down, GO_Biological_Process_2015, HMDB_Metabolites, KEGG_2016, NURSA_Human_Endogenous_Complexome, OMIM_Disease, PANTHER_2016, REACTOME_2016 and WikiPathways_2016.

Statistical analysis

All data analysis and visualization was completed using R, version 3.3.2 (R Core Team 2013).

Results

Integrated approach identifies genes in breast carcinogenesis

Combining information from multiple sources including text mining of PubMed and other biomedical libraries, a genomics data depository, a curated toxicogenomics database, and subject matter expert literature review, we created a gene panel focused on biological processes of breast development and breast carcinogenesis. This panel was built around 14 key concepts related to carcinogenesis and the mammary gland. Table 2 lists the five top-scoring genes for each characteristic in the BCScreen panel. Using the mammary characteristic as an example, the highest scoring genes include *BRCA1*, *BRCA2*, *PRLR*, *PRL* and *ERBB2*. *BRCA1* and *BRCA2* are well-studied tumor suppressor gene variants which are associated with a breast cancer risk five times higher than average [28,27]. The *PRL* and *PRLR* encodes the prolactin hormone and receptor respectively, and modulates the effects of prolactin in both normal and cancerous breast tissue. Lastly, the *ERBB2* gene encodes a member of the epidermal growth factor (EGF) receptor family of receptor tyrosine kinases known as HER2. HER2 plays an important role in cancer progression as an activator of signaling pathways that regulate cellular processes such as apoptosis and proliferation [59]. HER2 is overexpressed in 30% of breast tumors, as well as other cancer types [61]. The complete list of genes in BCScreen is included in Supplemental Table 2.

Intersection between BCScreen and other gene panels

Several gene sets have been curated to represent diverse biological mechanisms for studies on predictive and mechanistic toxicology. For example, due to the expense of whole transcriptome analysis, the Tox21 Working Group identified a subset of sentinel xenobiotic-responsive genes suitable for evaluating toxicity [31]. Called the S1500+, this gene list comprises approximately 2700 genes and was created using a hybrid data- and knowledge-driven approach to identify a gene set that encompassed key pathways in systemic toxicity.

Overall, we found that BCScreen encompassed a biological space not captured by other gene lists. We compared the S1500+ list to BCScreen, and found that less than half (230/500 or 46.0%) of BCScreen genes are included in the Tox21 S1500+. The overlap varied across characteristics (Table 2). For example, overlap with the S1500+ ranged from 21% for genes related to epigenetics to 82% for xenobiotic metabolism. Fewer than one-third of the genes involved in epigenetics, angiogenesis, growth hormones and immortalization are also on the S1500+. Fig. 3 shows summary scores for the 500 genes on BCScreen, and indicates whether the gene is included in the S1500+. In general, scores of genes that are included in S1500+ are similar to those that are only on BCScreen, although some high scoring genes involved in genotoxicity and xenobiotic metabolism are not included in S1500+. In some cases, such as for cell cycle, only lower scoring genes are missing from S1500+. Note that genes on BCScreen are assigned to a single characteristic but may have high relevance and scores for multiple characteristics (data not shown).

Three of the highest scoring genes in the BCScreen panel that do not appear on the S1500+gene list include *BRCA2*, *RAD51*, and *IL2*. *BRCA2* is a well-known tumor suppressor and breast cancer susceptibility gene. Its functions are essential for the repair of

DNA double strand breaks. *BRCA2* is most strongly associated with mammary characteristic, likely due to the large body of evidence on breast cancer susceptibility. Similarly, *RAD51* is a DNA repair gene with a central role homologous recombination repair of double strand breaks. This is consistent with its strong association and high ranking within the genotoxicity characteristic. Interestingly, *BRCA2* and *RAD51* are both central to repairing DNA double strand breaks and directly interact during this biological process. *RAD51* is overexpressed in different cancers, including breast cancer. The *IL2* gene encodes for a signaling molecule involved in key functions of the immune system. Overexpression of *IL2* has been observed in breast carcinoma biopsies, but not in normal breast tissues [57].

BCScreen shared some but not all genes with other panels. In a recent paper by Ryan et al. [42], a 47-gene biomarker for estrogen receptor α (ERα) activity was defined using microarray data from MCF-7 cells. Seven of these genes were included in BCScreen including *ALAD*, *AREG*, *CCND1*, *CXCL12*, *FHL2*, *PGR* and *RBBP8*. A set of 50 genes known as the PAM50 was developed to identify four breast cancer subtypes and predict recurrence. Of these 50 genes, 20 (40.0%) were included on BCScreen. Lastly, the Broad Institute's Library of Integrated Network-based Cellular Signatures (LINCS) compiled a list of "landmark" genes that are intended to represent the response of the full transcriptome in response to pharmacologic perturbation, known as the LINCS L1000. Of the 978 genes on the LINCS 1000, 84 (8.5%) were represented in BCScreen.

Enrichment analysis of BCScreen identifies key characteristics of breast carcinogenesis

Fig. 4 shows the top five significant pathways and terms from nine libraries which include WikiPathways, Reactome, Protein ANalysis THrough Evolutionary Relationships (PANTHER), Disease_perturbations_from_GEO_down (GEO), OMIM_Disease (OMIM), Kyoto Encyclopedia of Genes and Genomes (KEGG), Gene Ontology (GO) Biological Process, Human Metabolome Database (HMDB), and the Nuclear Receptor Signaling Atlas (NURSA). Note that enrichment is based on each gene set library, therefore the number of genes and annotated terms may be different for each library (see Supplemental Table 3). For these databases, the p-value was computed using the Fisher exact test and adjusted according to a z-score of the expected rank for each term [12]. Fig. 4 shows the absolute value of the magnitude of the log adjusted p-value for each pathway or process. For example, the p-value for the Sporadic Breast Cancer pathway identified in our list in the GEO library was 5.9×10^{-8}, graphed to the value of 8. Shading in Fig. 4 reports the extent of overlap between the genes identified as part of the Enrichr pathway and those in BC Screen. Using the Sporadic Breast Cancer pathway from GEO again, 35 BCScreen genes (9%) were represented in the 400 named as part of the Sporadic Breast Cancer Pathway in GEO. The two most significantly enriched pathways identified were the integrated pancreatic cancer pathway in WikiPathways, and the response to steroid hormones process in GO. The two pathways with the greatest overlap with BCScreen included resolution of D-loop structures (involved in DNA repair) in REACTOME and bladder cancer in KEGG.

The two disease-related libraries we probed (GEO and OMIM) showed enrichment for breast and other cancers. KEGG, PANTHER, WikiPathways and REACTOME all map genes to functions and processes in biological systems. Across these libraries, pathways related to multiple cancers, DNA damage and repair, and cell cycle pathways were most prominently represented in BCScreen. Interestingly, GO identified significant biological processes related to exposures associated with breast cancer including, responses to drugs, steroid hormones and alcohol. Metabolites most closely associated with the BCScreen panel included important steroid hormones and retinoic acid, a vitamin A metabolite important in growth and development. The NURSA database links nuclear receptors with their co-regulators, ligands, and downstream transcriptional targets. BCScreen showed enrichment of downstream targets involved in DNA repair and cell cycle regulation.

Sensitivity analysis

We purposefully selected genes with experimental evidence of altered expression in response to endocrine disruptors (GEO), and genes that had been identified as important to breast biology and cancer through ELR. We therefore assigned strong weights of 0.5 to each of those categories. An alternate approach to building the BCScreen panel would be to rely completely on the text mining of the biomedical literature to prioritize genes and assign them to the 14 characteristics of carcinogenesis. We evaluated how the genes obtained from the ELR and GEO MCF-7 data influenced the BCScreen gene list by comparing panels obtained if we did not consider those sources, i.e. if we gave them zero weight in the scoring. The extra 0.5 score for the 289 ELR genes increased the number of ELR genes on BCScreen from 76 to 276. The added 0.5 score for genes selected from GEO increased the number of GEO genes in BCScreen from 399 to 492, a small effect since such a large fraction of genes in the transcriptome were selected from the GEO query (13,256). Thus the ELR greatly influenced the final BCScreen panel. This contrast between the influence of weighting GEO versus ELR is not surprising because many more genes met the GEO selection criteria compared with the ELR criteria (13,256 versus 289). We also were interested in whether broader inclusion criteria for GEO would produce a substantially altered list. To this end, we created a supplemental version of BCScreen that used the unadjusted p-value to select differentially expressed genes in GEO. We found that 79% of genes were similar in adjusted and unadjusted lists (Supplemental Table 4).

Discussion

We developed and applied a novel gene prioritization framework to curate a panel of 500 genes responsive to chemicals that induce mammary toxicity and breast carcinogenesis. The purpose of this panel is to probe cellular or tissue responses to known mammary carcinogens and elucidate mechanisms relevant to breast cancer etiology. Gene expression profiles could be used to classify chemicals based on biological activity and highlight those whose profiles are similar to known breast or mammary carcinogens, thus strengthening chemical screening and prioritization strategies for breast cancer prevention.

Currently, there is little consensus on how gene prioritization or gene panel selection should be done. In a review on gene prioritization, Moreau and Trachevent [32], argue that computational approaches that integrate complex heterogeneous datasets offer a more thorough and unbiased assessment of candidate genes. Building on this concept, we employed both manual and automated gene curation and selection whereby data was extracted, weighted and combined from four separate data streams sources to create BCScreen. This modular approach provided flexibility, such that each source could be dynamically adjusted to modify the final list according to the priorities dictated by our central research question. This flexibility is central to this framework; investigators working on similar gene prioritization questions could tailor this approach to emphasize different carcinogenic characteristics, other disease endpoints, or highlight different functional genomic datasets. Furthermore, it allows researchers to add new data streams that incorporate other relevant genomic and toxicological data as it is generated, annotated and shared.

Fourteen characteristics relevant to carcinogenesis and breast cancer development identified from seminal papers formed the biological foundation for the BCScreen gene list [21,19,43,48]. These characteristics were each matched to one or more MeSH terms, and linked to a multi-source gene-MeSH association network. We used an NPMI association measure that ranks the degree of co-occurrence between each gene and the 14 key characteristics. Similar approaches have also used gene-MeSH annotations by generating a statistical test to determine whether a gene, author or disease has been over-represented within a particular MeSH term. For example, Cheung et al. [13] used the MeSH vocabulary to create MeSH Over-representation Profiles (MeSHOPs) as a means of identifying novel gene-disease relationships. Instead of using NPMI, their approach uses the Fisher exact test to determine whether a gene-topic association exceeded statistical significance or not, thus allowing for ranking based on p-value alone. However, our approach uses rank as a continuous variable, which allows us to more easily incorporate a scaled co-occurrence measure into a gene prioritization workflow.

For our investigation into breast cancer, we heavily weighted the panel such that genes selected through ELR and from functional genomic GEO data received the highest priority. The ELR was assigned a high score weighting because it had been manually curated by experts in breast cancer. The GEO dataset was assigned a similar high score weight so that genes that had been shown experimentally to respond to estrogen disruptors were included. Specifically, we selected a weight or score increment of 0.5 for the GEO and ELR datasets, which biased the list towards these data streams (see Sensitivity Analysis in Results). The allocation of 0.5 additional weight to the ELR list greatly altered the BCScreen panel by adding 200 genes. We will evaluate the value of information provided by the selected genes in future experimental work, including comparing the information conveyed by ELR versus the complete BCScreen. We expect the genes on BCScreen will evolve as experiments indicate which genes provide the most useful mechanistic information about breast carcinogens.

We used Enrichr to identify key pathways that are covered within BCScreen's 500 gene panel. The multiple enrichment analyses confirmed that cancer (in KEGG), breast cancer related pathways (in WikiPathways), and cell cycle processes (REACTOME and KEGG) were enriched in BCScreen. Breast cancer and other cancerswere also overrepresented in our gene panel (OMIM and GEO), as were processes including response to steroid hormones, apoptosis, and DNA repair. These results serve as a proof of principle that BCScreen captured pathways, metabolites and processes relevant to breast carcinogenesis.

Advances in next generation sequencing technologies have increased the efficiency of whole transcriptome analyses, however targeted RNA-sequencing using a gene panel like BCScreen has several advantages for identifying specific gene expression signatures in chemical screening. A targeted approach allows for the analysis of low abundance transcripts that might be lost in the "noise" of whole transcriptome analyses, potentially obscuring biologically important processes [51]. Secondly, a whole transcriptome approach to creating a mammary carcinogenesis biomarker would likely involve identifying a breast carcinogen gene expression signature by testing a large number of known mammary carcinogens. However, this approach is vulnerable to bias based on the set of chemicals used to establish the gene signature. In contrast, the gene prioritization framework used here was built based on an a priori designation of cancer-relevant pathways, not based on a single chemical or set of chemicals. Finally, while whole transcriptome research costs are decreasing, they remain significant; targeted gene panels allow resources to be allocated towards increasing numbers of replicates and chemicals tested, thus enhancing reproducibility and addressing important knowledge gaps about the toxicity of a vast number of unstudied chemicals.

BCScreen is shaped by a number of external factors that may introduce bias: the extent to which articles have been tagged with MeSH terms, how many have been annotated to specific genes, the number of different technologies that reproduce the same types of data, and publication bias. For example, as a newer area of research, genes linked to epigenetic studies may have comparatively lower rankings and fewer annotated genes, despite the fact that epigenetic mechanisms likely play a critical role in carcinogenesis. To address this, we utilized multiple input streams and built in modular flexibility to account for new research or updated sources. By allotting a predetermined number of genes to each characteristic, more novel areas of research like epigenetics are still represented in the panel and those genes did not need to compete with well-established pathways, such as angiogenesis. Nevertheless, future iterations of BCScreen could include a reallocation of genes to characteristics if experimental data demonstrate a measurable improvement.

BCScreen shares genes with other panels intended to capture gene expression changes relevant to breast cancer and general toxicity, but also includes many genes that are not in the other panels. For example, some genes overlap between BCScreen and PAM50 [54], the estrogenicity panel [42], and the Tox21 S1500+ panel [31]. This overlap indicates that while the other panels cover some of the biological space in BCScreen, our approach produced a unique gene set. One key difference between BCScreen and the PAM50 gene set is that PAM50 contains genes that have altered expression in cancerous cells. In contrast, we are more interested in genes whose expression in normal tissues will be disrupted by chemical exposure and genes that regulate biological processes that promote cancer development. As another example, Tox21's S1500+was created to represent the entire transcriptome in toxicological studies. We found that less than half of the BCScreen genes are in the S1500+ list, so we hypothesize that some important breast cancer pathways may not be represented. Experiments are planned to compare BCScreen, S1500+, and full transcriptomics as approaches for characterizing mechanisms of action for chemical-induced breast carcinogenesis. Comparing gene responses to breast carcinogens using these platforms will indicate which genes are most informative and likely will provide a basis for modifying the genes in the BCScreen panel. Overall this work will facilitate new discoveries about how chemicals cause breast cancer and help optimize standard toxicity testing approaches to be sensitive to these pathways.

There are limited data on which to base a decision about the optimum number of genes to include in this panel or to assign to a cancer characteristic. Our panel size of 500 genes is smaller than S1500+ (2700 genes) and LINCS L1000 (978 genes), both of which are designed to probe a complete set of known cellular response pathways. On the other hand, panels used for more specific and specialized purposes, such as the PAM50, tend to be much smaller. The S1500+ list, designed to identify systemic toxicity using 2700 genes, selected genes such that every key pathway included at least three genes. We allotted 33 genes for each of our characteristics to increase our confidence that each pathway or process would be adequately covered by the panel, and decided to assign an equal number of genes to each characteristic. We assigned a larger number of genes (71) to mammary as a characteristic to capture pathways specific to this tissue.

A next step for BCScreen is to use it in *in vivo* and *in vitro* toxicity studies to determine whether this panel captures critical mechanisms of chemical-induced breast carcinogenesis, and can identify new breast carcinogens. For example, probing BCScreen with breast carcinogens that have well-established mechanisms can indicate whether the BCScreen gene expression pattern accurately reflects that mechanism. Studies comparing gene expression signatures of known breast carcinogens with putative non-carcinogens and with carcinogens that don’t target the breast can begin to evaluate whether BCScreen can classify chemicals. A similar approach was reported by Ryan et al. [42], who developed a consensus gene expression signature for estrogen action, although our expectation is that breast carcinogens act by diverse biological pathways and we designed BCScreen to capture all of them. Another important area of work is to compare gene expression responses among different breast cell and tissue models, including standard cancer cell lines used in high throughput testing and realistic normal human breast tissue models.

In summary, recent advances in high throughput testing, the availability of functional genomic data, and machine learning algorithms like NPMI offer the opportunity to study complex networks of gene expression changes using tailored gene lists. In fact, these tools are already being integrated to derive hypotheses between genes, chemical exposures and disease [35]. However, given the large numbers of chemicals to be interrogated, and practical and cost limitations associated with whole transcriptome analysis, strategies to develop targeted gene panels are useful. The framework underlying BCScreen can serve as a useful model for other investigators, and as a research tool for those interested in the interaction between chemical exposures and breast cancer.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

The authors would like to thank L. Swetschinski for his valuable help with algorithm development and data processing. This work is supported by the California Breast Cancer Research Program (grant IDs 17UB-8708; DE-AC02-05CH11231; and 21-UB8100), NIEHS (1T32ES023769), Avon Foundation for Women (grant ID 02-2013-050), and Cedar Tree Foundation.

References

[1]. Ambrosone CB, Freudenheim JL, et al., Cigarette smoking, N-acetyltransferase 2 genetic polymorphisms, and breast cancer risk, J. Am. Med. Assoc. 276 (1996) 1494–1501.

[2]. American Cancer Society, Breast Cancer Facts & Figures 2013–2014, American Cancer Society, Inc., Atlanta, GA, 2014.

[3]. Apweiler R, Bairoch A, et al., UniProt: the Universal Protein knowledgebase, Nucl. Acids Res. 32 (Database issue) (2004) D115–119. [PubMed: 14681372]

[4]. Barrett T, Wilhite SE, et al., NCBI GEO: archive for functional genomics data sets-update, Nucl. Acids Res. 41 (Database issue) (2013) D991–995. [PubMed: 23193258]

[5]. Beam A, Rotroff D, et al. Quantitative Nuclease Protection Assays (qNPA) as Windows into Chemical-Induced Adaptive Response in Cultures of Primary Human Hepatocytes (Concentration and Time-Response), ToxCast Data Analysis Summit, Research Triangle Park, NC, 2009.

[6]. Benjamini Y, Drai D, et al., Controlling the false discovery rate in behavior genetics research. Behav. Brain Res. 125 (1–2) (2001) 279–284. [PubMed: 11682119]

[7]. Birnbaum LS, Fenton SE, Cancer and developmental exposure to endocrine disruptors, Environ. Health Perspect. Ill (4) (2003) 389–394.

[8]. Bouma G, Normalized (pointwise) mutual information in collocation extraction, in: Proceedings of the International Conference of the German Society for Computational Linguistics and Language Technology, 2009.

[9]. Brody J, Everyday exposures and breast cancer, Rev. Environ. Health 25 (1) (2010) 1–7. [PubMed: 20429152]

[10]. Cancer Genome Atlas Network, Comprehensive molecular portraits of human breast tumours, Nature 490 (7418) (2012) 61–70. [PubMed: 23000897]

[11]. Cardoso F, Vant Veer LJ, et al., 70-Gene signature as an aid to treatment decisions in early-stage breast cancer, New England J. Med. 375 (8) (2016) 717–729. [PubMed: 27557300]

[12]. Chen EY, Tan CM, et al., Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool, BMC Bioinf. 14 (2013) 128.

[13]. Cheung WA, Ouellette BF, et al., Quantitative biomedical annotation using medical subject heading over-representation profiles (MeSHOPs), BMC Bioinf. 13 (2012) 249.

[14]. Danaei G, Vander Hoorn S, et al., Causes of cancer in the world: comparative risk assessment of nine behavioural and environmental risk factors, Lancet 366 (9499) (2005) 1784–1793. [PubMed: 16298215]

[15]. Davis AP, Grondin CJ, et al., The Comparative Toxicogenomics Database: update 2017, Nucl. Acids Res. 45 (D1) (2017) D972–D978. [PubMed: 27651457]

[16]. Duan Q, Flynn C, et al., LINCS Canvas Browser: interactive web app to query, browse and interrogate LINCS L1000 gene expression signatures, Nucl. Acids Res. 42 (Web Server issue) (2014) W449–460. [PubMed: 24906883]

[17]. Edgar R, Domrachev M, et al., Gene Expression Omnibus: NCBI gene expression and hybridization array data repository, Nucl. Acids Res 30 (1) (2002) 207–210. [PubMed: 11752295]

[18]. Forman MR, Winn DM, et al., Environmental exposures, breast development and cancer risk: through the looking glass of breast cancer prevention, Reprod. Toxicol. 54 (2015) 6–10. [PubMed: 25499721]

[19]. Goodson WH 3rd, Lowe L, et al., Assessing the carcinogenic potential of low-dose exposures to chemical mixtures in the environment: the challenge ahead, Carcinogenesis 36 (Suppl 1) (2015) S254–S296. [PubMed: 26106142]

[20]. Hamilton JW, Kaltreider RC, et al., Molecular basis for effects of carcinogenic heavy metals on inducible gene expression, Environ. Health Perspect. 106 (Suppl 4) (1998) 1005–1015. [PubMed: 9703486]

[21]. Hanahan D, Weinberg RA, Hallmarks of cancer: the next generation, Cell 144 (5) (2011) 646–674. [PubMed: 21376230]

[22]. Hsu EL, Yoon D, et al., A proposed mechanism for the protective effect of dioxin against breast cancer, Toxicol. Sci. 98 (2) (2007) 436–444. [PubMed: 17517823]

[23]. IARC, A review of human carcinogens, Retrieved 7 4, 2014, from http://monographs.iarc.fr/ENG/Monographs/vol100A, 2012.

[24]. Interagency Breast Cancer & Environmental Research Coordinating Committee, Breast Cancer and the Environment: Prioritizing Prevention, 2013.

[25]. IOM, Breast Cancer and the Environment: A Life Course Approach, National Academies, Washington DC, 2011.

[26]. Karami F, Mehdipour P, A comprehensive focus on global spectrum of BRCA1 and BRCA2 mutations in breast cancer, Biomed Res. Int. (2013).

[27]. King MC, Marks JH, et al., Breast and ovarian cancer risks due to inherited mutations in BRCA1 and BRCA2, Science 302 (5645) (2003) 643–646. [PubMed: 14576434]

[28]. King MC, Marks JH, et al., Breast and ovarian cancer risk due to inherited mutations in BRCA1 & BRCA2, Science 302 (5645) (2003) 643–646. [PubMed: 14576434]

[29]. Kuleshov MV, Jones MR, et al., Enrichr: a comprehensive gene set enrichment analysis web server 2016 update, Nucl. Acids Res. 44 (W1) (2016) W90–W97. [PubMed: 27141961]

[30]. Latimer JJ, Johnson JM, et al., Nucleotide excision repair deficiency is intrinsic in sporadic stage I breast cancer, Proc. Natl. Acad. Sci. USA 107 (50) (2010) 21725–21730. [PubMed: 21118987]

[31]. Merrick BA, Paules RS, et al., Intersection of toxicogenomics and high throughput screening in the Tox21 program: an NIEHS perspective, Int. J. Biotechnol. 14 (1) (2015) 7–27. [PubMed: 27122658]

[32]. Moreau Y, Tranchevent LC, Computational tools for prioritizing candidate genes: boosting disease gene discovery, Nat. Rev. Genet. 13 (8) (2012) 523–536. [PubMed: 22751426]

[33]. Nebert DW, Dalton TP, The role of cytochrome P450 enzymes in endogenous signalling pathways and environmental carcinogenesis, Nat. Rev. Cancer 6 (12) (2006) 947–960. [PubMed: 17128211]

[34]. Palmer J, Rosenberg L, Cigarette smoking and risk of breast cancer, Epidemiol. Rev. 15 (1) (1993) 145–156. [PubMed: 8405197]

[35]. Patel CJ, Butte AJ, Predicting environmental chemical factors associated with disease-related gene expression data, BMC Med. Genomics 3 (2010) 17. [PubMed: 20459635]

[36]. Peck D, Crawford ED, et al., A method for high-throughput gene expression signature analysis, Genome Biol. 7 (7) (2006) R61. [PubMed: 16859521]

[37]. Pestalozzi BC, Tausch C, et al., Adjuvant treatment recommendations for patients with ER-positive/HER2-negative early breast cancer by Swiss tumor boards using the 21-gene recurrence score (SAKK 26/10), BMC Cancer 17 (1) (2017) 265. [PubMed: 28407750]

[38]. Rodgers KM, Udesky JO, et al., Environmental chemicals and breast cancer: An updated review of epidemiological literature informed by biological mechanisms, Environ. Res. 160 (2017) 152–182. [PubMed: 28987728]

[39]. Rudel RA, Ackerman JM, et al., New exposure biomarkers as tools for breast cancer epidemiology, biomonitoring, and prevention: a systematic approach based on animal evidence, Environ. Health Perspect. 112 (9) (2014).

[40]. Rudel RA, Attfield KR, et al., Chemicals causing mammary gland tumors in animals signal new directions for epidemiology, chemicals testing, and risk assessment for breast cancer prevention, Cancer 109 (12 Suppl) (2007) 2635–2666. [PubMed: 17503434]

[41]. Rudel RA, Fenton SE, et al., Environmental exposures and mammary gland development: state of the science, public health implications, and research recommendations, Environ. Health Perspect. 119 (8) (2011) 1053–1061. [PubMed: 21697028]

[42]. Ryan N, Chorley B, et al., Moving toward integrating gene expression profiling into high-throughput testing: a gene expression biomarker accurately predicts estrogen receptor alpha modulation in a microarray compendium, Toxicol. Sci. 151 (1) (2016) 88–103. [PubMed: 26865669]

[43]. Schwarzman MR, Ackerman JM, et al., Screening for chemical contributions to breast cancer risk: a case study for chemical safety evaluation, Environ. Health Perspect. (2015).

[44]. Shioda T, Chesnes J, et al., Importance of dosage standardization for interpreting transcriptomal signature profiles: evidence from studies of xenoestrogens, Proc. Natl. Acad. Sci. USA 103 (32) (2006) 12033–12038. [PubMed: 16882715]

[45]. Shioda T, Rosenthal NF, et al., Expressomal approach for comprehensive analysis and visualization of ligand sensitivities of xenoestrogen responsive genes, PNAS 110 (41) (2013) 16508–16513. [PubMed: 24062438]

[46]. Singletary KW, Gapstur SM, Alcohol and breast cancer: review of epidemiologic and experimental evidence and potential mechanisms, JAMA 286 (17) (2001) 2143–2151. [PubMed: 11694156]

[47]. Skinner MK, Endocrine disruptors in 2015: epigenetic transgenerational inheritance, Nat. Rev. Endocrinol. 12 (2) (2016) 68–70. [PubMed: 26585656]

[48]. Smith MT, Guyton KZ, et al., Key characteristics of carcinogens as a basis for organizing data on mechanisms of carcinogenesis, Environ. Health Perspect. 124 (6) (2016) 713–721. [PubMed: 26600562]

[49]. Spink DC, Wu SJ, et al., Induction of CYP1A1 and CYP1B1 by benzo(k)fluor-anthene and benzo(a)pyrene in T-47D human breast cancer cells: roles of PAH interactions and PAH metabolites, Toxicol. Appl. Pharmacol. 226 (3) (2008) 213–224. [PubMed: 17919675]

[50]. Suzek BE, Wang Y, et al., UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches, Bioinformatics 31 (6) (2015) 926–932. [PubMed: 25398609]

[51]. Tarazona S, Garcia-Alcalde F, et al., Differential expression in RNA-seq: a matter of depth, Genome Res. 21 (12) (2011) 2213–2223. [PubMed: 21903743]

[52]. Tranchevent LC, Ardeshirdavani A, et al., Candidate gene prioritization with Endeavour, Nucl. Acids Res. 44 (W1) (2016) W117–W121. [PubMed: 27131783]

[53]. Trombino AF, Near RI, et al., Expression of the aryl hydrocarbon receptor/transcription factor (AhR) and AhR-regulated CYP1 gene transcripts in a rat model of mammary tumorigenesis, Breast Cancer Res. Treat. 63 (2) (2000) 117–131. [PubMed: 11097088]

[54]. Wallden B, Storhoff J, et al., Development and verification of the PAM50-based Prosigna breast cancer gene signature assay, BMC Med. Genomics 8 (2015) 54. [PubMed: 26297356]

[55]. Watford SM, Grashow R, et al., Novel application of normalized pointwise mutual information (NPMI) to mine biomedical literature for chemical links to breast carcinogenesis, In preparation, 2017.

[56]. Williams JA, Phillips DH, Mammary expression of xenobiotic metabolizing enzymes and their potential role in breast cancer, Cancer Res. 60 (17) (2000) 4667–4677. [PubMed: 10987265]

[57]. Esquivel-Velázquez M, Ostoa-Saloma P, et al., The role of cytokines in breast cancer development and progression, J. Interferon Cytokine Res. 35 (1) (2015) 1–16. [PubMed: 25068787]

[58]. Kleinstreuer NC, Dix DJ, et al., In vitro perturbations of targets in cancer hallmark processes predict rodent chemical carcinogenesis, Toxicol. Sci. 131 (1) (2013) 40–55. [PubMed: 23024176]

[59]. Liu P, Cheng H, et al., Targeting the phosphoinositide 3-kinase (PI3K) pathway in cancer, Nat. Rev. Drug Discov. 8 (2011) 627–644.

[60]. Russo J, Russo IH, Molecular Basis of Breast Cancer: Prevention and Treatment, Springer-Verlag, New York, 2004.

[61]. Tan M, Yu D, Molecular mechanisms of erbB2-mediated breast cancer chemoresistance, Adv. Exp. Med. Biol. 608 (2007) 119–129. [PubMed: 17993237]

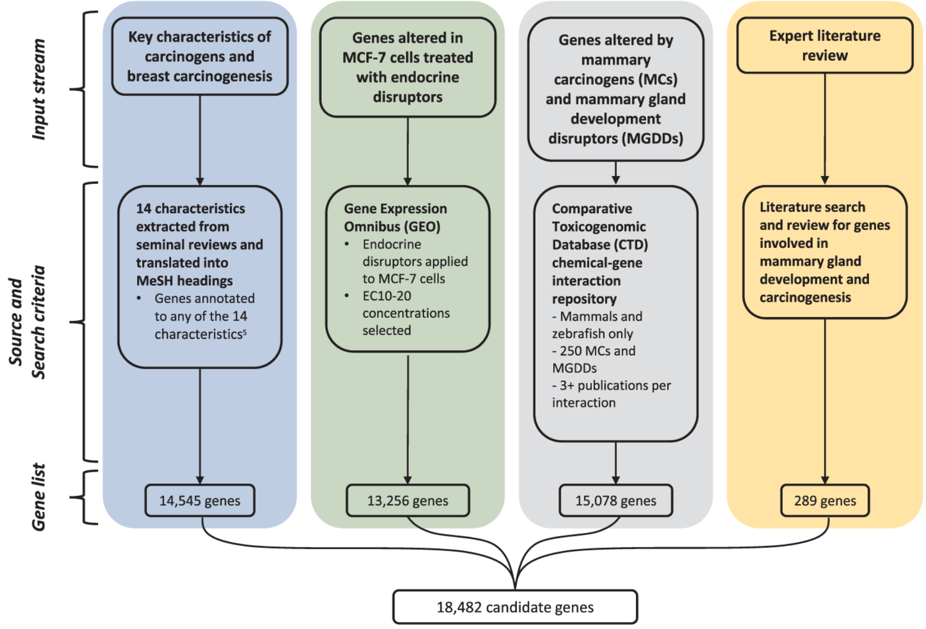
Input stream
Source and Search criteria
Gene list
Key characteristics of carcinogens and breast carcinogenesis
Genes altered in MCF-7 cells treated with endocrine disruptors
Genes altered by mammary carcinogens (MCs) and mammary gland development disruptors (MGDDs)
Expert literature review
14 characteristics extracted from seminal reviews and translated into MeSH headings
• Genes annotated to any of the 14 characteristics5
Gene Expression Omnibus (GEO)
• Endocrine disruptors applied to MCF-7 cells
• EC10-20 concentrations selected
Comparative Toxicogenomic Database (CTD) chemical-gene interaction repository
- Mammals and zebrafish only
- 250 MCs and MGDDs
- 3+ publications per interaction
Literature search and review for genes involved in mammary gland development and carcinogenesis
14,545 genes
13,256 genes
15,078 genes
289 genes
18,482 candidate genes

Fig. 1.

Conceptual data inputs and selection criteria for candidate gene list. Legend: MeSH: Medical Subject Headings; NPMI: normalized pointwise mutual information; MCF-7: Michigan Cancer Foundation-7 mammary tissue cell line; EC: effective concentration; MC: mammary carcinogens; MGDD: mammary gland developmental disruptors.

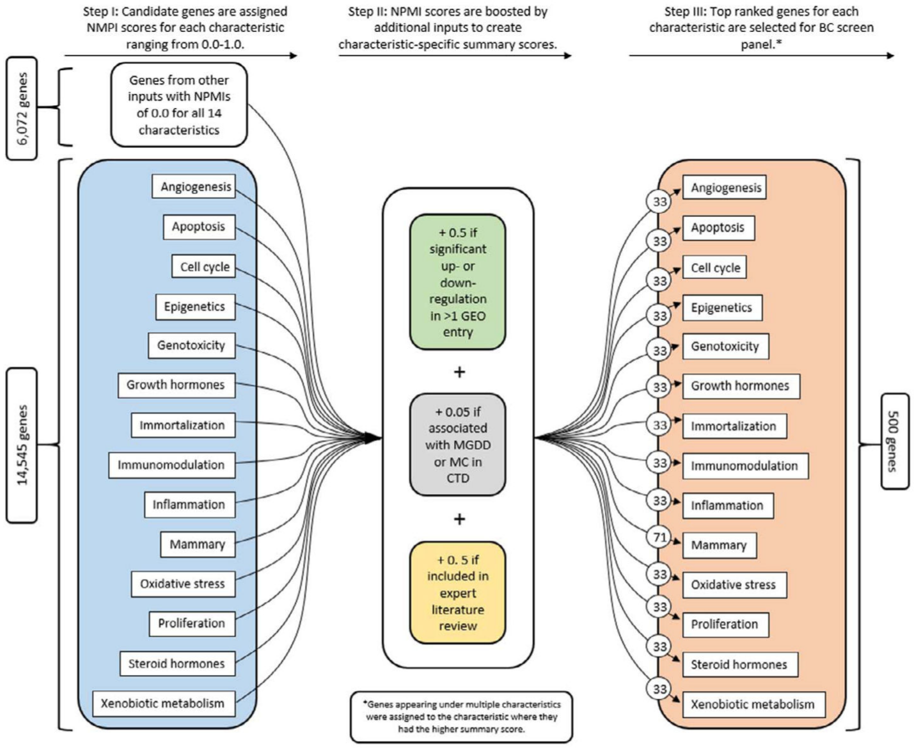
Step I: Candidate genes are assigned NMPI scores for each characteristic ranging from 0.0-1.0.
Step II: NPMI scores are boosted by additional inputs to create characteristic-specific summary scores.
Step III: Top ranked genes for each characteristic are selected for BC screen panel.*
6,072 genes
Genes from other inputs with NPMIs of 0.0 for all 14 characteristics
14,545 genes
Angiogenesis
Apoptosis
Cell cycle
Epigenetics
Genotoxicity
Growth hormones
Immortalization
Immunomodulation
Inflammation
Mammary
Oxidative stress
Proliferation
Steroid hormones
Xenobiotic metabolism
+ 0.5 if significant up- or down-regulation in >1 GEO entry
+
+ 0.05 if associated with MGDD or MC in CTD
+
+ 0. 5 if included in expert literature review
33
33
33
33
33
33
33
33
33
71
33
33
33
33
Angiogenesis
Apoptosis
Cell cycle
Epigenetics
Genotoxicity
Growth hormones
Immortalization
Immunomodulation
Inflammation
Mammary
Oxidative stress
Proliferation
Steroid hormones
Xenobiotic metabolism
500 genes
*Genes appearing under multiple characteristics were assigned to the characteristic where they had the higher summary score.

Fig. 2.

Ranking, scoring and selection of genes for final BCScreen panel.

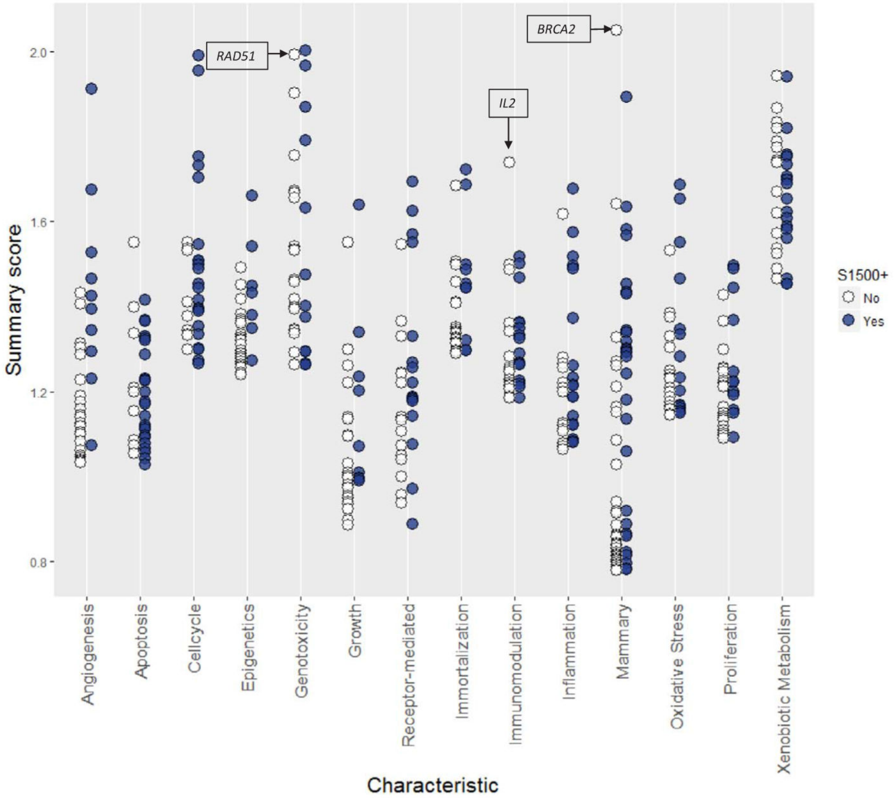

Fig. 3.

Summary scores for 500 BCScreen genes assigned to 14 characteristics. Blue circles indicate genes that appear in both BCScreen and the S1500+. Empty circles represent genes included in BCScreen only.

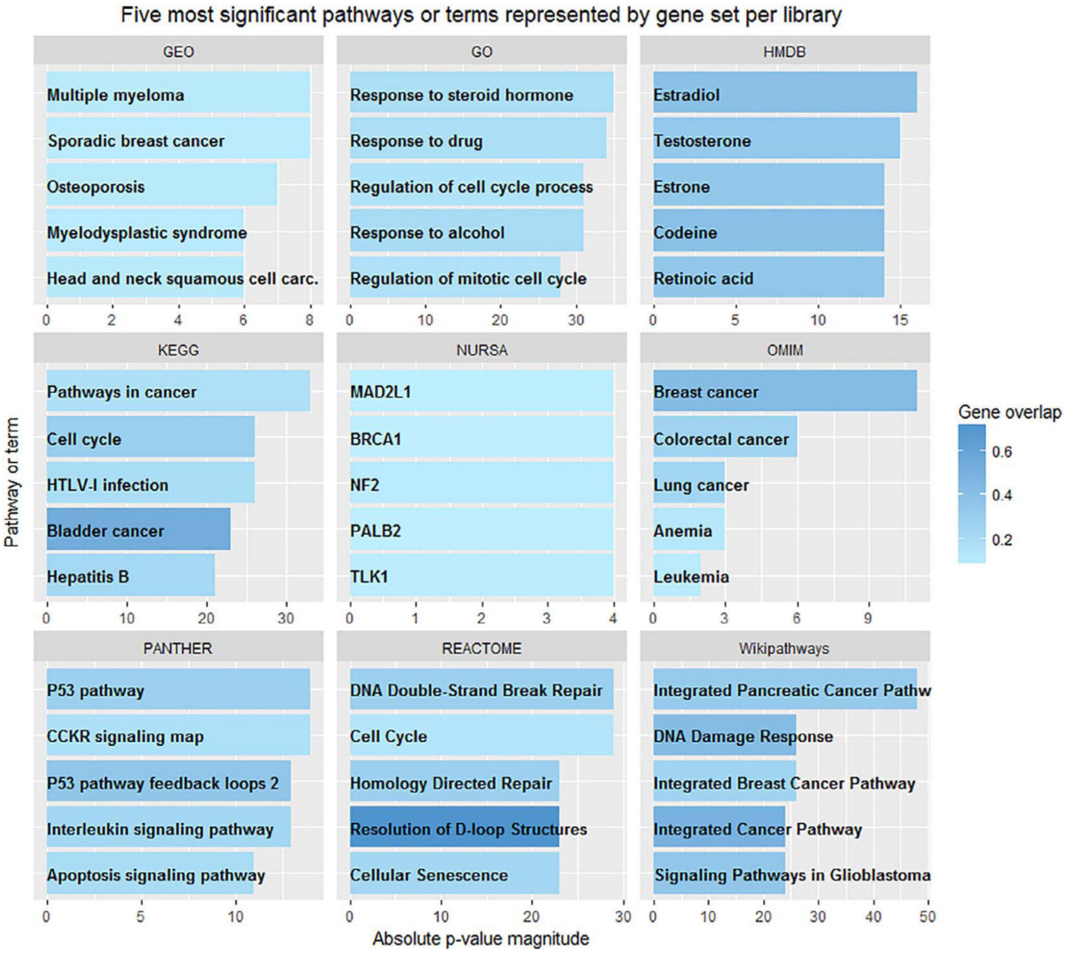

Fig. 4.

Significantly enriched pathways, metabolites and processes in nine Enrichr libraries. Five most significant results are presented as ordered by absolute log p-value magnitude. Shading indicates the percentage of BCScreen genes represented in each pathway or term. Legend: HTLV-1: Human T-lymphotrophic virus-1; DM: Diabetes mellitus; CCKR: Gastrin and cholecystokinin receptor; care.: carcinoma.

Table 1

Keywords, sources and Medical Subject Heading (MeSH) terms for key carcinogenesis characteristics.

Characteristic	Description	Hanahan and Weinberg [21]	Smith et al. [48]	Schwarzman et al. [43]	Goodson et. al. [19]	MeSH term
Angiogenesis	Blood vessel formation and growth	✓		✓	✓	Angiogenesis, pathologic
Apoptosis	Controlled cell death	✓	✓	✓	✓	Angiogenesis, physiologic Apoptosis
Cell cycle	Events and processes related to cell division and duplication		✓	✓		Cell cycle
Epigenetics	Mechanisms that modulate and regulate gene expression without altering underlying DNA sequences		✓	✓	✓	Epigenomics
Genotoxicity	Damage to genetic material (e.g. DNA, RNA)	✓	✓	✓	✓	DNA damage
Growth hormones	Endogenous internal signaling systems related to growth		✓	✓	✓	DNA repair Growth hormone
Immortalization	Evasion of normal celluar senescence and uncontrolled division	✓	✓	✓	✓	Cell survival
Immunomodulation	Regulatory adjustment of the immune system	✓	✓	✓	✓	Immune system
Inflammation	Mechanisms in response to infection or tissue injury	✓	✓	✓	✓	Inflammation
Mammary	Anatomical structures or processes related to human female breats or other mammalian milk-secreting organs			✓		Breast
Oxidative stress	Production and detoxificataion of reactive oxygen species		✓	✓		Oxidative stress
Proliferation	Cell population growth	✓	✓	✓	✓	Cell proliferation
Steroid hormones	Cholesterol derived hormones that control reproduction, metabolism, the immune system		✓	✓		Gonadal steroid hormones
Xenobiotic metabolism	Metabolic pathways that biotransform exogenous compounds		✓	✓		Xenobiotics

Table 2

Top scoring genes and percentage inclusion in S1500+ by characteristic.

Characteristic	Genes with highest summary score	Percentage of genes appearing on S1500+
Angiogenesis	*VEGFA*, *THBS1*, *HIF1A*, *MMP2*, *TWIST1*	30.3%
Apoptosis	*PTRH2*, *CAV1*, *MMP11*, *SRC*, *TP53*	75.8%
Cell cycle	*CCNB1*, *CDK1*, *CDK2*, *CDK4*, *CCNA1*	69.7%
Epigenetics	*GREB1*, *HMGCS2*, *TFF3*, *TSSC4*, *JUNB*	21.2%
Genotoxicity	*EXO1*, *ATM*, *XRCC5*, *RAD51*, *BRIP1*	39.4%
Growth hormones	*IGF1*, *GH2*, *GH1*, *GHR*, *CSH2*	30.3%
Immortalization	*BCL2L1*, *BCL2*, *ZNF165*, *ARMC1*, *BIRC5*	30.3%
Immunomodulation	*IL2*, *CD69*, *GATA3*, *CD38*, *CSF1*	48.5%
Inflammation	*IL6*, *PYCARD*, *IL1B*, *F3*, *IL8*	54.5%
Mammary	*BRCA2*, *BRCA1*, *PRLR*, *PRL*, *ERBB2*	39.4%
Oxidative stress	*NOL3*, *HMOX1*, *CAT*, *ENO1*, *HSPA8*	42.4%
Proliferation	*NOTCH1*, *MYC*, *CTNNB1*, *TBX3*, *CCND2*	36.7%
Steroid hormones	*ESR1*, *ESR2*, *PGR*, *CYP17A1*, *STS*	51.5%
Xenobiotic metabolism	*NR1I3*, *NR1I2*, *EPHX1*, *NAT1*, *CYP1A2*	81.8%