GeneSigDB: A Manually Curated Database and Resource for Analysis of Gene Expression Signatures

doi:10.1093/nar/gkr901

GeneSigDB: a manually curated database and resource for analysis of gene expression signatures

Aedín C. Culhane[1,2,*], Markus S. Schröder[1], Razvan Sultana[1], Shaita C. Picard[1], Enzo N. Martinelli[1], Caroline Kelly[1], Benjamin Haibe-Kains[1,2], Misha Kapushesky[3], Anne-Alyssa St Pierre[1], William Flahive[1], Kermshlise C. Picard[1], Daniel Gusenleitner[1], Gerald Papenhausen[1], Niall O'Connor[1], Mick Correll[1] and John Quackenbush[1,3,4,*]

[1]Biostatistics and Computational Biology, Dana-Farber Cancer Institute, 450 Brookline Avenue, Boston, MA 02215, [2]Biostatistics, Harvard School of Public Health, 677 Huntington Avenue, Boston, MA 02115, USA, [3]EMBL Outstation- Hinxton, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, UK and [4]Cancer Biology, Dana-Farber Cancer Institute, 450 Brookline Avenue, Boston, MA 02215, USA

Received September 19, 2011; Accepted October 5, 2011

ABSTRACT

GeneSigDB (http://www.genesigdb.org or http://compbio.dfci.harvard.edu/genesigdb/) is a database of gene signatures that have been extracted and manually curated from the published literature. It provides a standardized resource of published prognostic, diagnostic and other gene signatures of cancer and related disease to the community so they can compare the predictive power of gene signatures or use these in gene set enrichment analysis. Since GeneSigDB release 1.0, we have expanded from 575 to 3515 gene signatures, which were collected and transcribed from 1604 published articles largely focused on gene expression in cancer, stem cells, immune cells, development and lung disease. We have made substantial upgrades to the GeneSigDB website to improve accessibility and usability, including adding a tag cloud browse function, facetted navigation and a 'basket' feature to store genes or gene signatures of interest. Users can analyze GeneSigDB gene signatures, or upload their own gene list, to identify gene signatures with significant gene overlap and results can be viewed on a dynamic editable heatmap that can be downloaded as a publication quality image. All data in GeneSigDB can be downloaded in numerous formats including .gmt file format for gene set enrichment analysis or as a R/Bioconductor data file. GeneSigDB is available from http://www.genesigdb.org.

INTRODUCTION

Accurately curated and annotated gene sets have emerged as essential tools for the analysis of large, complex biological datasets. Gene set analysis (GSA) is widely used in the analysis and interpretation of gene expression profiling data (1–4), evolutionary relationships (5), genomic associations—including QTL analysis (6), genotyping (7) and SNP chips (8)—and even for cross platform integration of genomics data (9). GSA aims to find sets of genes that collectively distinguish two phenotypes, even if the genes in the set are not significantly different when tested individually. This reflects the fact that genes within the cell function as members of complex networks and pathways, often with multiple, overlapping functions. As a result, direct comparisons of genes may miss biologically important connections that are only seen when these related genes are assessed collectively.

Gene sets have also become invaluable tools for characterizing and distinguishing phenotypic states. In breast cancer, for example, several gene expression signatures have been developed as commercial diagnostic assays (10) and new methods are being developed that combine the predictive strength of multiple gene signatures to increase their prognostic power (11).

Gene set resources can be broadly divided into those which assign a gene to collections based on 'known' gene or protein interactions or functional activity and those that include gene lists from high-throughput experimental assays. Functional and pathway databases such as Gene Ontology (GO), KEGG and Reactome capture published descriptions of cellular pathways and gene functions (12), including, in the case of GO, functional predictions inferred from orthologous sequences (13). However, these

*To whom correspondence should be addressed. Tel: +1 617 632 2468; Fax: +1 617 582 7760; Email: aedin@jimmy.harvard.edu
Correspondence may also be addressed to John Quackenbush. Tel: +1 617 582 8163; Fax: +1 617 582 7760; Email: johnq@jimmy.harvard.edu

resources are incomplete as we have not yet been able to comprehensively and completely catalog the functions of all genes in the genome (13).

High-throughput experiments, such as microarray expression profiling and RNA-seq have also produced large numbers of potentially informative gene lists. Most genomics papers present one or more gene signatures that reportedly correlate with experimental phenotypes. While there has been some controversy over the value of individual gene sets, due to the fact that many fail to fully replicate in independent data sets, the analysis of the collected gene lists defined for similar phenotypes has been demonstrated to provide meaningful biological insight (14).

Despite tremendous interest in using gene signatures, public repositories such as GEO and ArrayExpress (15,16) store primary gene expression data but fail to capture the gene sets that are the end product of published analyses. Without a systematic way of reporting these, the gene sets often appear only in published tables or figures or in supplementary materials hosted on the author's or the journal's website. And as there are no accepted standards for reporting gene sets, they often appear with non-standard gene identifiers, making comparison to other lists, or even to the original data, a significant challenge. Because of these limitations, gene sets from published research studies are often inaccessible to automated computational analysis.

In August 2009, we created GeneSigDB (17) as a repository for gene sets that had been systematically collected and manually curated from published articles indexed by PubMed. Our approach in building GeneSigDB was to capture gene signatures from the literature as published, to map them to standard identifiers using transparent, reproducible protocols and to freely provide these to the research community together with some elementary analytical tools. Since its launch, GeneSigDB had 7918 web hits with 4404 hits in 2010 and 3354 so far this year, suggesting that this resource is of value to the biomedical research community.

GROWTH OF GENESIGDB DATABASE

GeneSigDB has grown considerably since its introduction (Figure 1), nearly doubling in size with each subsequent release. GeneSigDB 4.0, released in September 2011, contains 3515 human, mouse and rat gene sets curated from 1604 published articles. While we have continued to focus on gene sets related to cancer and stem cells, we now also include signatures for development, inflammation and immune regulation, and lung disease and we have begun to catalog signatures for miRNA expression and proteomics. The content of GeneSigDB and its composition are summarized in Tables 1 and 2.

GeneSigDB had minimal overlap with other gene signatures resources when we compared the overlap of publications curated by MSigDB (18) and CCancer (19) to GeneSigDB. Only 198/1604 (12%) publications were curated by both GeneSigDB (n = 1604) and MSigDB (n = 786, Release 3.0). GeneSigDB and MSigDB are both manually curated, but CCancer gene lists are computationally extracted from publications in 100 journals indexed by PubMed (19), and ~30% of publications in CCancer are also manually curated in GeneSigDB. To estimate of curation quality, we examined 121 publications that were curated by all three gene signatures resources. The number of gene signatures identified in these 121 publications were 428 287 and 123 by MSigDB, GeneSigDB and CCancer respectively. MSigDB and GeneSigDB captured more data per publication that the automated curation of CCancer.

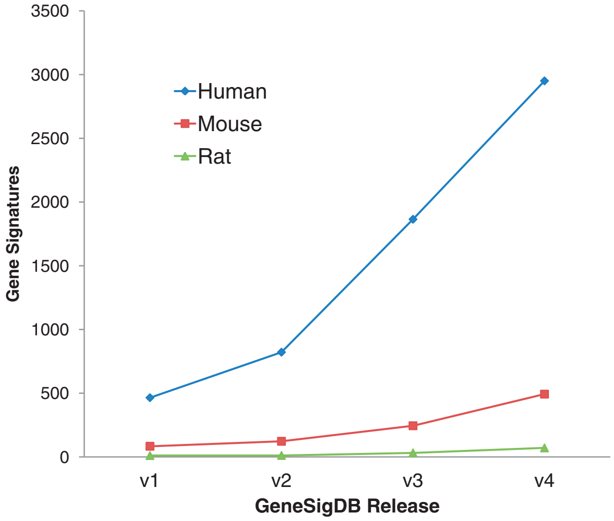


Figure 1. Growth of GeneSigDB. GeneSigDB has grown considerably over 4 database releases (August 2009, March 2010, December 2010, September 2011). The most recent release (Release 4.0, September 2011) contains 3515 human, mouse and rat gene sets curated from 1604 published articles.

Table 1. Number of processed articles and extracted gene signatures (by species) in GeneSigDB

	Human	Mouse	Rat	Total
Gene Signatures	2951	493	71	3515
Publications (PMIDs)	1368	208	39	1604*
Genes (EnsEMBL gene IDs)	20 478	16 009	5110	

*There were 10 articles with human and mouse gene signatures, and 1 article with human and rat gene signatures.

GENE SIGNATURE COLLECTION AND CURATION PIPELINE

The primary data objects in GeneSigDB are genes, published articles and gene signatures from those publications. We define a gene signature as a set of gene identifiers that were experimentally derived from analysis of gene, protein or miRNA expression.

Published articles likely to contain gene signature are identified using predefined PubMed searches as described at http://compbio.dfci.harvard.edu/genesigdb/documentation.jsp. We download and read each article, identify

Table 2. Most common disease MeSH terms associated with articles in GeneSigDB[a]

MeSH Terms	Publications
Breast neoplasms	248
Lung neoplasms	97
Prostatic neoplasms	73
Disease progression	69
Neoplasm metastasis	66
Ovarian neoplasms	66
Adenocarcinoma	65
Cell transformation, neoplastic	62
Neoplasm invasiveness	62
Carcinoma, squamous cell	58
Liver neoplasms	56
Carcinoma, hepatocellular	51
Lymphatic metastasis	42
Colonic neoplasms	38
Neoplasms	37
Precursor cell lymphoblastic leukemia–lymphoma	37
Stomach neoplasms	35
Neovascularization, pathologic	34
Disease models, animal	33
Genetic predisposition to disease	32
Pancreatic neoplasms	30
Chromosome aberrations	29
Carcinoma	28
Leukemia, myeloid, acute	28
Brain neoplasms	27
Carcinoma, non-small-cell lung	27
Leukemia, myeloid	25
Neoplasm recurrence, local	25
Leukemia, lymphocytic, chronic, B-cell	24

[a] Ranking of Diseases MeSH Terms (MeSH prefix code category C) associated with 1552 publications in GeneSigDB. A total of 63 publications were not annotated with MeSH Terms. More details are provided in documentation on the GeneSigDB website.

tables or figures containing gene signatures, and transcribe these from the main body of the article, its supplementary materials or websites referenced within the article (Figure 2). We manually transcribe the entire table and then use a pipeline based on Biomart (20) to map the published gene identifiers to EnsEMBL IDs, creating standardized gene sets.

The number of genes in the standardized sets may not be equal to that reported in the source publication and this may occur for a variety of reasons. Gene identifiers reported in the article maybe have been retired, reported probes may now be recognized as non-specific or they may map to multiple genes, or gene identifiers maybe invalid due to inaccurate reporting or to MS Excel gene name conversion errors (21). GeneSigDB users have the option of seeing either the original or standardized versions of a gene set and in the current release version users can compare the original and standardized lists side-by-side. In addition, we have improved the visibility of unmapped genes so that users can identify unmapped genes in a GeneSigDB standardized table

Gene signatures in the database are identified by unique SigIDs and SigNames. The SigIDs combine the PubMed ID of the paper from which the signature was derived and the table or figure in which it was reported. For example, SigID 11823860-SuppTable2 refers to a gene signature obtained from Table 2 in the supplementary material from an article with PMID 11823860. The SigName is designed to be a more descriptive, human-readable identifier; the SigName associated with 11823860-SuppTable2 is Breast_van'tVeer02_231genes_PoorPrognosisSignature, which indicates it is a signature of poor prognosis in breast

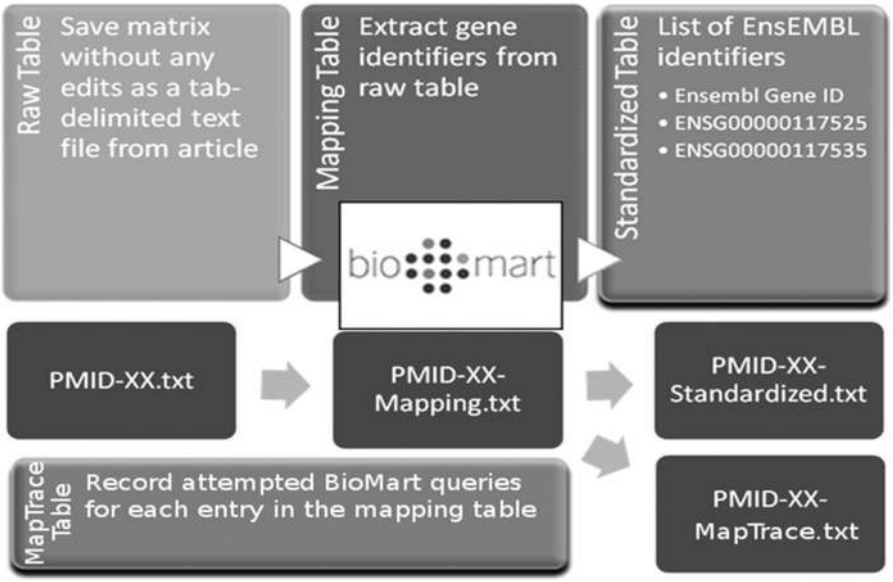

Raw Table
Save matrix without any edits as a tab-delimited text file from article
Mapping Table
Extract gene identifiers from raw table
Standardized Table
List of EnsEMBL identifiers
• Ensembl Gene ID
• ENSG00000117525
• ENSG00000117535
bio mart
PMID-XX.txt
PMID-XX-Mapping.txt
PMID-XX-Standardized.txt
MapTrace Table
Record attempted BioMart queries for each entry in the mapping table
PMID-XX-MapTrace.txt

Figure 2. Overview of the GeneSigDB Data Curation pipeline. Gene signatures in tables or figures are transcribed from published articles indexed in PubMed and we then use a pipeline based on Biomart (20) to map all published gene identifiers to EnsEMBL IDs, to create standardized gene sets.

cancer, contains 231 genes and was published by van 't Veer and colleagues in 2002.

BROWSING AND SEARCHING GeneSigDB

We provide two search tools for finding signatures in the database, one based on publications and the other based on genes. The publication search tool allows users to enter one or more search terms, such as author name, article title, journal name or keywords (such as disease type), and these are then searched against the full text of articles represented in the database to find those best meeting the search criteria. In release 4.0 we added the ability to search Medical Subject Headings (MeSH) terms associated with each publication.

Users can also search for genes and their annotated properties, including gene names and synonyms, functional classifications such as GO terms, InterPro domains, KEGG or Reactome Pathways or almost any valid gene identifiers including gene symbol, Entrez gene ID, EnsEMBL gene ID, RefSeq ID or common commercial microarray probe IDs.

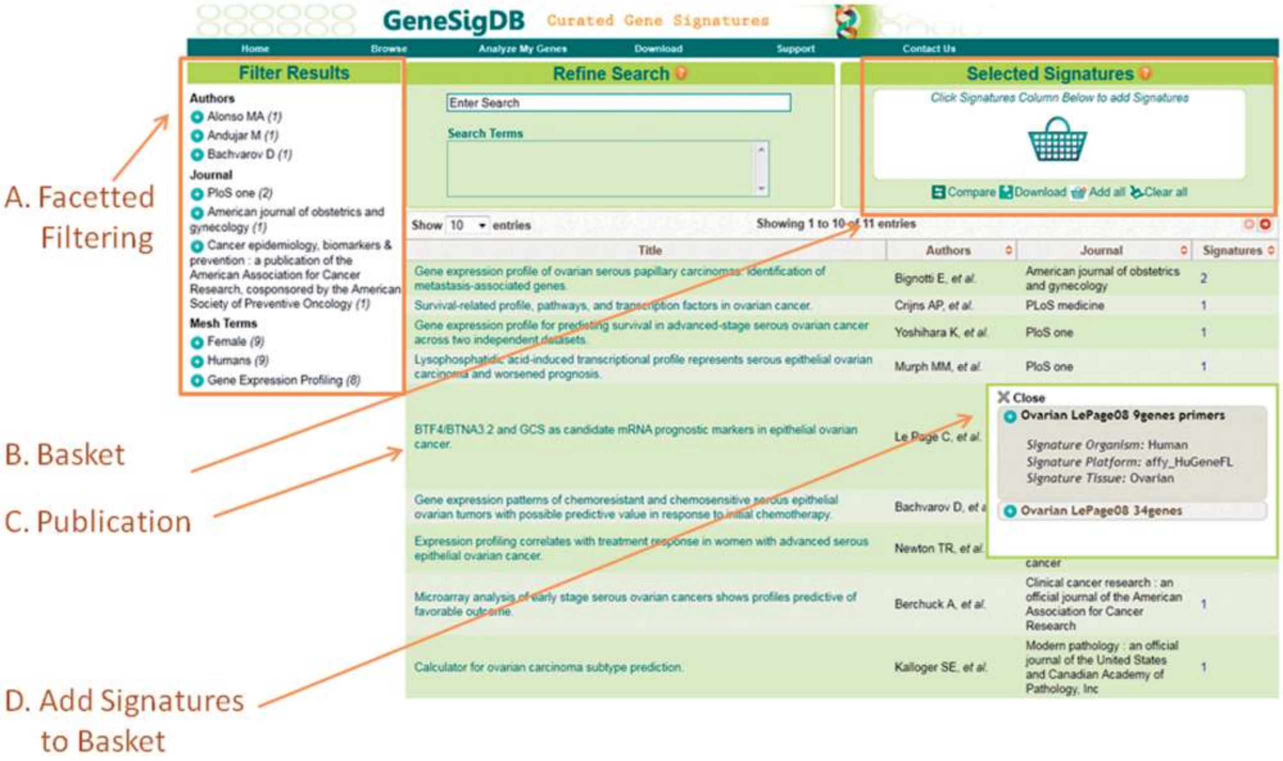
Results can be further refined by applying additional search criteria or using faceted terms associated with publication or gene search results (Figure 3). For either gene-based or publication-based searches, users can collect their results and load these into a 'Shopping Basket', allowing the signatures to be collected signatures using a variety of independent criteria and to then viewed, downloaded or compared (Figure 3).

DATA VIEWS

Clicking on a publication, gene or gene signature will open up a data type-specific view for each of these. The publication view provides information about the published article, its authors, an abstract and a list of gene signatures associated with that publication. The gene view provides annotation on a gene and a list of gene signatures which contain that gene. Each of these pages includes links to one or more gene signatures. The gene signature view now provides a dynamic table presenting both the original and standardized gene lists, each of which can be used to sort and filter the signature.

COMPARING GENE SIGNATURES

One common question is whether a particular gene set overlaps with others that have been reported. For published gene lists, we have pre-computed the pairwise similarity between all gene signatures using a one-tailed Fisher's exact test (which is equivalent to a hypergeometric distribution test) with *P*-values corrected for multiple testing. When a user clicks 'Related Signatures' from a gene signature view, or 'Compare' gene signatures in the Shopping Basket the most similar signatures are presented both in list and graphical form (Figure 4).



Authors
- Alonso MA (1)
- Andujar M (1)
- Bachvarov D (1)

Journal
- PloS one (2)
- American journal of obstetrics and gynecology (1)
- Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology (1)

Mesh Terms
- Female (9)
- Humans (9)
- Gene Expression Profiling (8)

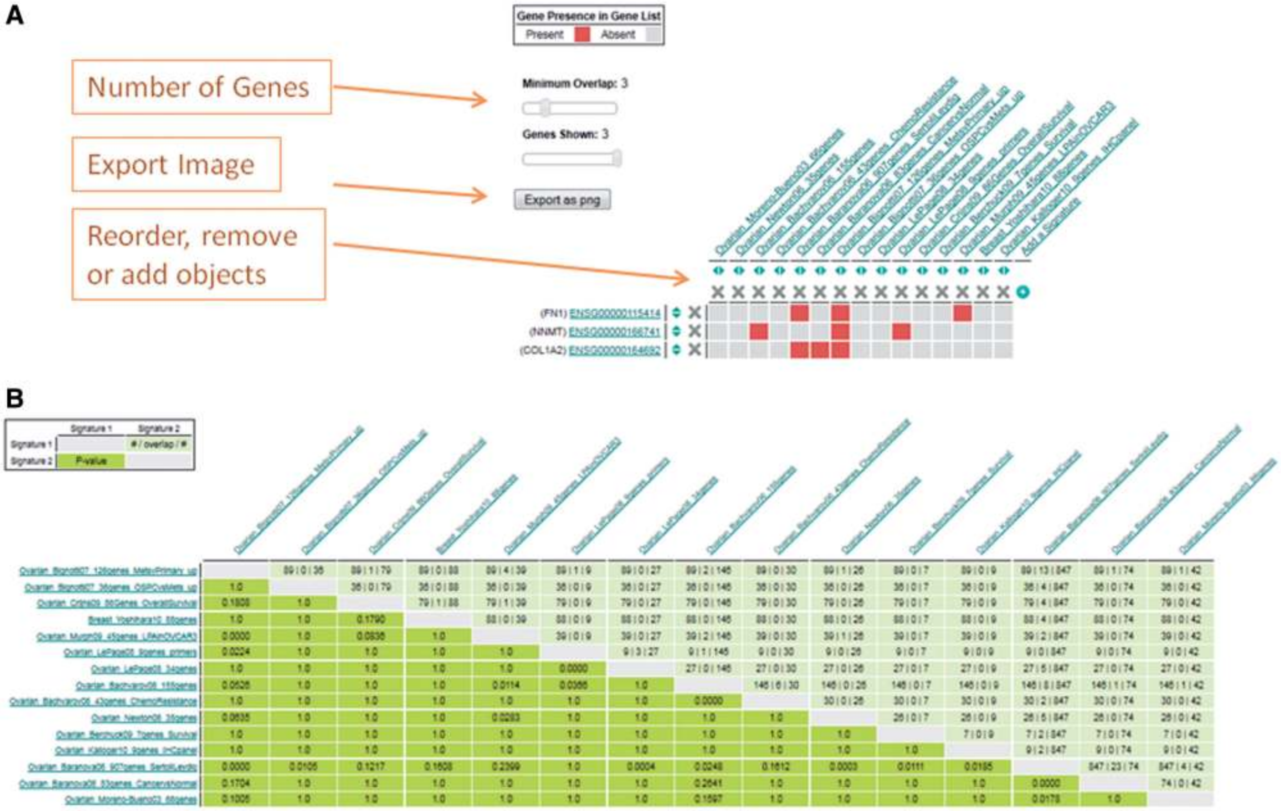
Show 10 entries Showing 1 to 10 of 11 entries

Title	Authors	Journal	Signatures
Gene expression profile of ovarian serous papillary carcinomas: identification of metastasis-associated genes.	Bignotti E, *et al.*	American journal of obstetrics and gynecology	2
Survival-related profile, pathways, and transcription factors in ovarian cancer.	Crijns AP, *et al.*	PLoS medicine	1
Gene expression profile for predicting survival in advanced-stage serous ovarian cancer across two independent datasets.	Yoshihara K, *et al.*	PloS one	1
Lysophosphatidic acid-induced transcriptional profile represents serous epithelial ovarian carcinoma and worsened prognosis.	Murph MM, *et al.*	PloS one	1
BTF4/BTNA3.2 and GCS as candidate mRNA prognostic markers in epithelial ovarian cancer.	Le Page C, *et al.*		
Gene expression patterns of chemoresistant and chemosensitive serous epithelial ovarian tumors with possible predictive value in response to initial chemotherapy.	Bachvarov D, *et a*		
Expression profiling correlates with treatment response in women with advanced serous epithelial ovarian cancer.	Newton TR, *et al.*	cancer	
Microarray analysis of early stage serous ovarian cancers shows profiles predictive of favorable outcome.	Berchuck A, *et al.*	Clinical cancer research : an official journal of the American Association for Cancer Research	1
Calculator for ovarian carcinoma subtype prediction.	Kalloger SE, *et al.*	Modern pathology : an official journal of the United States and Canadian Academy of Pathology, Inc	1

- Ovarian LePage08 9genes primers

  Signature Organism: Human
  Signature Platform: affy_HuGeneFL
  Signature Tissue: Ovarian
- Ovarian LePage08 34genes

Figure 3. Screenshot showing the (**A**) faceting and the (**B**) 'Shopping Basket' search features when we performed a publication search for 'serous ovarian cancer' that returned a list of 11 publications. Further detail about (**C**) publications or (**D**) Signature can be viewed by clicking on their respective links. By default 10 results are shown but up to 100 search results can be viewed. Selecting 'Add All' will add the 15 gene signatures associated with the 11 displayed publications to the basket so they can be compared or downloaded.



Gene Presence in Gene List			
Present	🟥	Absent	▨

Minimum Overlap: 3

Genes Shown: 3

Export as png

	Ovarian_Moreno-Bueno03_66genes	Ovarian_Newton06_35genes	Ovarian_Bachvarov06_155genes	Ovarian_Bachvarov06_42genes_ChemoResistance	Ovarian_Baranova06_907genes_SertoliLeydig	Ovarian_Baranova06_83genes_CancervsNormal	Ovarian_Bignotti07_126genes_MetsPrimary_up	Ovarian_Bignotti07_36genes_OSPCvsMets_up	Ovarian_LePage08_34genes	Ovarian_LePage08_9genes_primers	Ovarian_Crijns09_86Genes_OverallSurvival	Ovarian_Berchuck09_7genes_Survival	Ovarian_Murph09_45genes_LPAinOVCAR3	Breast_Yoshihara10_88genes	Ovarian_Kalloger10_9genes_IHCpanel	Add a Signature
(FN1) ENSG00000115414	▨	▨	▨	▨	🟥	▨	🟥	▨	▨	▨	▨	▨	🟥	▨	▨	
(NNMT) ENSG00000166741	▨	▨	🟥	▨	▨	▨	🟥	▨	▨	🟥	▨	▨	▨	▨	▨	
(COL1A2) ENSG00000164692	▨	▨	▨	▨	🟥	🟥	🟥	▨	▨	▨	▨	▨	▨	▨	▨	

	Signature 1	Signature 2
Signature 1		# / overlap / #
Signature 2	P-value	

	Ovarian Bignotti07 126genes MetsPrimary up	Ovarian Bignotti07 36genes OSPCvsMets up	Ovarian Crijns09 86Genes OverallSurvival	Breast Yoshihara10 88genes	Ovarian Murph09 45genes LPAinOVCAR3	Ovarian LePage08 9genes primers	Ovarian LePage08 34genes	Ovarian Bachvarov06 155genes	Ovarian Bachvarov06 42genes ChemoResistance	Ovarian Newton06 35genes	Ovarian Berchuck09 7genes Survival	Ovarian Kalloger10 9genes IHCpanel	Ovarian Baranova06 907genes SertoliLeydig	Ovarian Baranova06 83genes CancervsNormal	Ovarian Moreno-Bueno03 66genes
Ovarian Bignotti07 126genes MetsPrimary up		89|0|36	89|1|79	89|0|88	89|4|39	89|1|9	89|0|27	89|2|146	89|0|30	89|1|26	89|0|17	89|0|9	89|13|847	89|1|74	89|1|42
Ovarian Bignotti07 36genes OSPCvsMets up	1.0		36|0|79	36|0|88	36|0|39	36|0|9	36|0|27	36|0|146	36|0|30	36|0|26	36|0|17	36|0|9	36|4|847	36|0|74	36|0|42
Ovarian Crijns09 86Genes OverallSurvival	0.1808	1.0		79|1|88	79|1|39	79|0|9	79|0|27	79|0|146	79|0|30	79|0|26	79|0|17	79|0|9	79|4|847	79|0|74	79|0|42
Breast Yoshihara10 88genes	1.0	1.0	0.1790		88|0|39	88|0|9	88|0|27	88|0|146	88|0|30	88|0|26	88|0|17	88|0|9	88|4|847	88|0|74	88|0|42
Ovarian Murph09 45genes LPAinOVCAR3	0.0000	1.0	0.0836	1.0		39|0|9	39|0|27	39|2|146	39|0|30	39|1|26	39|0|17	39|0|9	39|2|847	39|0|74	39|0|42
Ovarian LePage08 9genes primers	0.0224	1.0	1.0	1.0	1.0		9|3|27	9|1|146	9|0|30	9|0|26	9|0|17	9|0|9	9|0|847	9|0|74	9|0|42
Ovarian LePage08 34genes	1.0	1.0	1.0	1.0	1.0	0.0000		27|0|146	27|0|30	27|0|26	27|0|17	27|0|9	27|5|847	27|0|74	27|0|42
Ovarian Bachvarov06 155genes	0.0526	1.0	1.0	1.0	0.0114	0.0366	1.0		146|6|30	146|0|26	146|0|17	146|0|9	146|8|847	146|1|74	146|1|42
Ovarian Bachvarov06 42genes ChemoResistance	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.0000		30|0|26	30|0|17	30|0|9	30|2|847	30|0|74	30|0|42
Ovarian Newton06 35genes	0.0635	1.0	1.0	1.0	0.0283	1.0	1.0	1.0	1.0		26|0|17	26|0|9	26|5|847	26|0|74	26|0|42
Ovarian Berchuck09 7genes Survival	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0		7|0|9	7|2|847	7|0|74	7|0|42
Ovarian Kalloger10 9genes IHCpanel	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0		9|2|847	9|0|74	9|0|42
Ovarian Baranova06 907genes SertoliLeydig	0.0000	0.0105	0.1217	0.1608	0.2399	1.0	0.0004	0.0248	0.1612	0.0003	0.0111	0.0185		847|23|74	847|4|42
Ovarian Baranova06 83genes CancervsNormal	0.1704	1.0	1.0	1.0	1.0	1.0	1.0	0.2641	1.0	1.0	1.0	1.0	0.0000		74|0|42
Ovarian Moreno-Bueno03 66genes	0.1005	1.0	1.0	1.0	1.0	1.0	1.0	0.1597	1.0	1.0	1.0	1.0	0.0178	1.0	

Figure 4. Screen shots of the 'Analyze My Gene Lists' and 'Comparison View'. The gene overlap between 15 serous ovarian cancer gene signatures selected in Figure 3 were analyzed. (**A**) shows the gene overlap between signature, where presence and absence of a gene are indicates by a red or gray pixel in the heatmap. The image can be edited, reordered or genes and gene signatures can be added or removed before the image is exported as a publication quality image. (**B**) shows results of a Fisher's exact test of enrichment between the 15 gene signatures. These results can also be visualized as a list.

A heatmap is used to visualize the overlaps of genes in a selection of gene signatures. In the heatmap, gray pixels indicate no overlap and red indicates two signatures sharing a common gene. Users can reorder, add or remove genes or gene signatures from the heatmap or use a sliding selector to define a minimum number of genes required for overlap. Users can also edit the heatmap column and row labels and export the heatmap as a publication quality image.

Those wishing to compare their own gene lists to the published lists in GeneSigDB can select 'Analyze My Genes' from the top menu bar. This brings up an interface that allows them to paste in their own gene list or to upload a file containing a gene list. If the gene list is not a list of EnsEMBL IDs, it is converted using BioMart, tested for overlap with the gene sets in the database, and reported using the table and heatmap views described above.

DOWNLOADING GENESIGDB DATA

As described above, users may perform a search to create a custom selection of gene signatures in their basket and these can be downloaded as a compressed file. In addition, all of the data in GeneSigDB is freely available for download. Users selecting 'Download' from the top menu bar are taken to a page where they can download the current release and previous versions of the database. The data are available in a variety of formats, including a tab-delimited flat file, GSEA gmt format and as an R/Bioconductor RData file.

PROGRAMMATIC ACCESS TO GENESIGDB DATA

Release 4.0 of GeneSigDB includes expanded programmatic access to the database through a Java RESTful web service. We use the reference implementation of JAX-RS found at Glassfish (Jersey: https://jersey.dev.java.net) to provide the REST HTTP functionality and use the Glassfish reference implementation of JAXB (JAXB: https://jaxb.dev.java.net), for the XML transformation. GeneSigDB provides REST services to retrieve each of the major objects in GeneSigDB (GeneSignature, Gene and Publication) along with all of their ancillary member objects. These objects are in either XML or JSON format. The REST request is made over HTTP by creating a URL with an embedded key that will then GET the specified resource. The Accepts portion of the HTTP request header will determine the MIME (and format) of the response. Further details and examples of how these queries should be constructed are available in GeneSigDB online documentation.

ADDING LINK OUT RESOURCES TO ADD VALUE TO GENESIGDB

In building GeneSigDB, we chose to report the published signatures rather than attempt to re-analyze the original data described in each of the manuscripts we extract from PubMed. Not only is re-analysis technically challenging because of incomplete metadata, but there are other projects that attempt to do this including Oncomine, Exalt (22) the Gene Expression Atlas (GXA) project (23) and our OncoSurf project, which is focused on finding signatures that predict survival (http://cccb.dfci.harvard.edu/oncosurf/).

The most comprehensive resource is GXA (http://www.ebi.ac.uk/gxa) whose developers have reanalyzed over 5500 gene expression and RNA-sequencing studies to identify expression profiles in 19 000 cellular or clinical phenotypes. In partnership with GXA, we are providing 'link out' access to GXA, providing visualization of the GeneSigDB gene sets so that users can easily see which genes within a signature are significantly associated with specific phenotypes in GXA.

SUBMITTING DATA FOR PUBLICATION IN GENESIGDB

Although we have been accepting submission to GeneSigDB by email, Release 4.0 includes a web form for signature submission. Users can also use this form to suggest updates to gene signatures currently in GeneSigDB.

USER CASES STUDIES

Although GeneSigDB is a relatively new database, it has already been used to advance our understanding of cancer and disease in new and interesting ways. Abba and colleagues used GeneSigDB 2.0 to retrieve breast cancer gene signatures ($n = 42$) to identify the 117 most common genes across those signatures. They found the common genes to be enriched for those associated with response to steroid hormone stimulus, and the cell cycle. Their meta-signature of the 42 GeneSigDB gene signatures was capable of predicting overall survival ($P < 0.0001$) and relapse-free survival ($P < 0.0001$) in patients with early-stage breast carcinoma. GeneSigDB has also been used to develop methods for Transcription factor binding site analysis (24) and graph theory algorithms (25)

ARCHITECTURE OF WEB INTERFACE TO GENESIGDB

The web interface to GeneSigDB (http://compbio.dfci.harvard.edu/genesigdb) is based on HTML, CSS, Javascript, JSP, XML and Java 1.6 technologies. The application runs on an Apache Tomcat 6 web application server running on a CentOS 5 Linux server. Front-end interactivity makes use of the jQuery 1.4+ Javascript library and server-side processing is based on the Apache Solr Enterprise Search Server 1.4. GeneSigDB is based on a hybrid approach between traditional database technologies and Solr indexes to catalog and organize the data. Search functions are performed using the Solr server to create a high-performance search engine. Other site functions are often performed using a database backend. Both the database and the indexes contain the same information, but they are organized in different ways to take advantage of the relative strengths of each technology.

SUMMARY AND FUTURE DIRECTIONS

GeneSigDB addresses the important need within the community to standardize gene expression signatures so they can easily be compared to each other and used in GSA. The current release represents an almost 8-fold increase in the number of gene sets from the first version and includes 3515 gene signatures derived from the analysis of cancer, stem cells, immune system function, development and lung disease; at present it is the largest source of cancer-specific gene signatures available. Based on the needs of our users, we have made considerable efforts to increase the functionality of the GeneSigDB website to facilitate mining and analysis of gene signatures.

In the future, we hope to expand GeneSigDB functionality to provide links so that users can analyze connections between genes using our predictive networks application (http://www.predictivenetworks.org) and test whether a gene signature is prognostic or associated with a known SNP using OncoSurf (http://cccb.dfci.harvard.edu/oncosurf).

Although there have been other attempts to catalog gene sets, we believe that GeneSigDB represents a significant advance in both the quantity of gene signature data we have amassed and the quality of the analysis we perform. By standardizing both the way we refer to these gene sets and the manner in which they are mapped to standard formats, we have created an infrastructure that can be scaled and extended to capture the growing number of genomic profiles that are being created and published. In a time when we as a community have become aware of the need for reproducible research, such standardization is essential to assure that the results of genomic studies can be broadly used and replicated in independent analysis. Further, the availability of standardized signatures creates an opportunity for us to more fully leverage the prior knowledge that has been gained by expert analysis of individual studies so that we can more rapidly advance our understanding of the nature of a broad range of human diseases.

ACKNOWLEDGEMENTS

We thank Dr Oliver Hofmann, Prof. Winston Hide and Dr Levi Waldron for their useful suggestions, which have helped us to improve GeneSigDB.

FUNDING

This work was supported by National Institutes for Health National Library of Medicine (1R01 LM010129), National Cancer Institute (1U19 CA148065), Genome

Research Institute (1P50 HG004233); Dana-Farber Cancer Institute Women's Cancers Program (to A.C.C); Claudia Adams Barr foundation.

Conflict of interest statement. None declared.

REFERENCES

1. Subramanian,A., Tamayo,P., Mootha,V.K., Mukherjee,S., Ebert,B.L., Gillette,M.A., Paulovich,A., Pomeroy,S.L., Golub,T.R., Lander,E.S. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA*, **102**, 15545–15550.
2. Mootha,V.K., Lindgren,C.M., Eriksson,K.-F., Subramanian,A., Sihag,S., Lehar,J., Puigserver,P., Carlsson,E., Ridderstråle,M., Laurila,E. *et al.* (2003) PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat. Genet*, **34**, 267–273.
3. Goeman,J.J. and Bühlmann,P. (2007) Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics*, **23**, 980–987.
4. Liberzon,A., Subramanian,A., Pinchback,R., Thorvaldsdóttir,H., Tamayo,P. and Mesirov,J.P. (2011) Molecular signatures database (MSigDB) 3.0. *Bioinformatics*, **27**, 1739–1740.
5. Bardelli,A. and Velculescu,V.E. (2005) Mutational analysis of gene families in human cancer. *Curr. Opin. Genet. Dev*, **15**, 5–12.
6. Al-Shahrour,F., Arbiza,L., Dopazo,H., Huerta-Cepas,J., Mínguez,P., Montaner,D. and Dopazo,J. (2007) From genes to functional classes in the study of biological systems. *BMC Bioinformatics*, **8**, 114.
7. Wu,C., Delano,D.L., Mitro,N., Su,S.V., Janes,J., McClurg,P., Batalov,S., Welch,G.L., Zhang,J., Orth,A.P. *et al.* (2008) Gene set enrichment in eQTL data identifies novel annotations and pathway regulators. *PLoS Genet*, **4**, e1000070.
8. Raychaudhuri,S., Plenge,R.M., Rossin,E.J., Ng,A.C.Y., Purcell,S.M., Sklar,P., Scolnick,E.M., Xavier,R.J., Altshuler,D. and Daly,M.J. (2009) Identifying relationships among genomic disease regions: predicting genes at pathogenic SNP associations and rare deletions. *PLoS Genet*, **5**, e1000534.
9. Montaner,D. and Dopazo,J. Multidimensional Gene Set Analysis of Genomic Data. *PLoS One*, **5**, e.0010348.
10. Perou,C.M. and Børresen-Dale,A.-L. (2011) Systems biology and genomics of breast cancer. *Cold Spring Harb. Perspect. Biol.*, **3**, t.a003293.
11. Zhao,X., Rødland,E.A., Sørlie,T., Naume,B., Langerød,A., Frigessi,A., Kristensen,V.N., Børresen-Dale,A.-L. and Lingjærde,O.C. (2011) Combining gene signatures improves prediction of breast cancer survival. *PLoS ONE*, **6**, e17845.
12. Croft,D., O'Kelly,G., Wu,G., Haw,R., Gillespie,M., Matthews,L., Caudy,M., Garapati,P., Gopinath,G., Jassal,B. *et al.* (2011) Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Res.*, **39**, D691–D697.
13. The Reference Genome Group of the Gene Ontology Consortium. (2009) The Gene Ontology's Reference Genome Project: a unified framework for functional annotation across species. *PLoS Comput. Biol.*, **5**, e1000431.
14. Fan,C., Prat,A., Parker,J.S., Liu,Y., Carey,L.A., Troester,M.A. and Perou,C.M. (2011) Building prognostic models for breast cancer patients using clinical variables and hundreds of gene expression signatures. *BMC Med. Genomics*, **4**, 3.
15. Parkinson,H., Kapushesky,M., Kolesnikov,N., Rustici,G., Shojatalab,M., Abeygunawardena,N., Berube,H., Dylag,M., Emam,I., Farne,A. *et al.* (2009) ArrayExpress update–from an archive of functional genomics experiments to the atlas of gene expression. *Nucleic Acids Res.*, **37**, D868–D872.
16. Barrett,T., Troup,D.B., Wilhite,S.E., Ledoux,P., Evangelista,C., Kim,I.F., Tomashevsky,M., Marshall,K.A., Phillippy,K.H., Sherman,P.M. *et al.* (2010) NCBI GEO: archive for functional genomics data sets—10 years on. *Nucleic Acids Res.*, **39**, D1005–D1010.
17. Culhane,A.C., Schwarzl,T., Sultana,R., Picard,K.C., Picard,S.C., Lu,T.H., Franklin,K.R., French,S.J., Papenhausen,G., Correll,M. *et al.* (2010) GeneSigDB–a curated database of gene expression signatures. *Nucleic Acids Res.*, **38**, D716–D725.
18. Liberzon,A., Subramanian,A., Pinchback,R., Thorvaldsdóttir,H., Tamayo,P. and Mesirov,J.P. (2011) Molecular signatures database (MSigDB) 3.0. *Bioinformatics*, **27**, 1739–1740.
19. Dietmann,S., Lee,W., Wong,P., Rodchenkov,I. and Antonov,A.V. (2010) CCancer: a bird's eye view on gene lists reported in cancer-related studies. *Nucleic Acids Res.*, **38**, W118–W123.
20. Haider,S., Ballester,B., Smedley,D., Zhang,J., Rice,P. and Kasprzyk,A. (2009) BioMart Central Portal—unified access to biological data. *Nucleic Acids Res.*, **37**, W23–W27.
21. Zeeberg,B.R., Riss,J., Kane,D.W., Bussey,K.J., Uchio,E., Linehan,W.M., Barrett,J.C. and Weinstein,J.N. (2004) Mistaken identifiers: gene name errors can be introduced inadvertently when using Excel in bioinformatics. *BMC Bioinformatics*, **5**, 80.
22. Wu,J., Qiu,Q., Xie,L., Fullerton,J., Yu,J., Shyr,Y., George,A.L. and Yi,Y. (2009) Web-based interrogation of gene expression signatures using EXALT. *BMC Bioinformatics*, **10**, 420.
23. Kapushesky,M., Emam,I., Holloway,E., Kurnosov,P., Zorin,A., Malone,J., Rustici,G., Williams,E., Parkinson,H. and Brazma,A. (2010) Gene expression atlas at the European bioinformatics institute. *Nucleic Acids Res.*, **38**, D690–D698.
24. Van de Sande,B., Atak,Z.K. and Aerts,S. (2011) Discovery of regulators for co-expressed human genes using large sequence search spaces. *Bioinformatics for Regulatory Genomics (BioRegSIG) Special Interest Group (SIG) at ISMB*. http://light.ece.ohio.edu/bioreg/2011/pages/vandesande.pdf. (16 July 2011, date last accessed).
25. Georgii,E., Tsuda,K. and Schölkopf,B. (2011) Multi-way set enumeration in weight tensors. *Machine Learning*, **82**, 123–155.