SPECIAL SERIES: INFORMATICS TOOLS FOR CANCER RESEARCH AND CARE

original reports abstract

OncoMX: A Knowledgebase for Exploring Cancer Biomarkers in the Context of Related Cancer and Healthy Data

Hayley M. Dingerdissen, MS[1]; Frederic Bastian, PhD[2,3]; K. Vijay-Shanker, PhD[4]; Marc Robinson-Rechavi, PhD[2,3]; Amanda Bell, MS[1]; Nikhita Gogate, MSc[1]; Samir Gupta, MS[4]; Evan Holmes, BS[1]; Robel Kahsay, PhD[1]; Jonathon Keeney, PhD[1]; Heather Kincaid, BS[5]; Charles Hadley King, MS[1]; David Liu, BS[5]; Daniel J. Crichton, MS[5]; and Raja Mazumder, PhD[1]

PURPOSE The purpose of OncoMX[1] knowledgebase development was to integrate cancer biomarker and relevant data types into a meta-portal, enabling the research of cancer biomarkers side by side with other pertinent multidimensional data types.

METHODS Cancer mutation, cancer differential expression, cancer expression specificity, healthy gene expression from human and mouse, literature mining for cancer mutation and cancer expression, and biomarker data were integrated, unified by relevant biomedical ontologies, and subjected to rule-based automated quality control before ingestion into the database.

RESULTS OncoMX provides integrated data encompassing more than 1,000 unique biomarker entries (939 from the Early Detection Research Network [EDRN] and 96 from the US Food and Drug Administration) mapped to 20,576 genes that have either mutation or differential expression in cancer. Sentences reporting mutation or differential expression in cancer were extracted from more than 40,000 publications, and healthy gene expression data with samples mapped to organs are available for both human genes and their mouse orthologs.

CONCLUSION OncoMX has prioritized user feedback as a means of guiding development priorities. By mapping to and integrating data from several cancer genomics resources, it is hoped that OncoMX will foster a dynamic engagement between bioinformaticians and cancer biomarker researchers. This engagement should culminate in a community resource that substantially improves the ability and efficiency of exploring cancer biomarker data and related multidimensional data.

JCO Clin Cancer Inform 4:210-220. © 2020 by American Society of Clinical Oncology

ASSOCIATED CONTENT

Data Supplement

Author affiliations and support information (if applicable) appear at the end of this article.

Accepted on February 5, 2020 and published at ascopubs.org/journal/cci on March 6, 2020: DOI https://doi.org/10.1200/CCI.19.00117

INTRODUCTION

Cancer biomarkers are molecules that can be assayed from bodily fluids or tissues whose presence indicates some process(es) associated with cancer.[2] Molecular characterization of cancers can lead to pan-cancer or cancer type–specific biomarkers, which are becoming integral components of risk assessment, pathologic diagnosis, monitoring of disease progression, and therapeutic decisions.[3,4] This is especially true when biomarker assays enable the identification of subpopulations amenable to treatment by targeted molecular therapies. Cancer biomarkers are commonly individual genes or proteins, multigene/protein panels, or biomolecules, like glycans, analyzed from urine, blood, stool, or other biologic sample. Biomarker diversity is ever growing and may also include findings from image analyses,[5] gut microbiome abundance,[6] and others.

Attempts at molecular characterization of cancer samples have increased as a result of advances in high-throughput sequencing and other technologic progress.[7,8] Paired with increased funding opportunities, such as those enabled through the Precision Medicine Initiative of 2015, it is not surprising to find increased reports of potentially novel cancer biomarkers.[9] However, reproducibility of initial findings, clinical validation, and access to harmonized biomarker and associated data remain formidable challenges.[7]

While issues of reproducibility and clinical validation must be addressed by biomarker research and regulatory communities, usability and accessibility of findings can be addressed in the public domain. In fact, a number of research groups already make cancer biomarker data available, including EDRN—the Early Detection Research Network focusing on the research and development of biomarkers and technologies for the clinical application of early cancer detection strategies[10]; CBD—the Colorectal Cancer Biomarker Database containing colorectal cancer biomarkers reported from articles in PubMed;[11]

CONTEXT

Key objective:

To develop a data model and central, unified, integrated Web resource to enable improved exploration of cancer biomarkers in the context of related evidence.

Knowledge generated:

A data model was developed to describe cancer biomarker evidence and that was capable of accommodating heterogeneously structured extant data and extensible to diverse new data types. Data integrated and unified through this model were made available through OncoMX, a knowledgebase and Web portal for exploring cancer biomarker data and related evidence.

Relevance:

Cancer biomarker evidence can be accumulated across various resources from a single access point through OncoMX to generate summary reports for a given gene, including available information on clinical status and relevance. Furthermore, OncoMX is linking to several other resources, improving the reach of each resource reciprocally and increasing the utility of any given data set.

ResMarkerDB—the database of biomarkers of drug response to monoclonal antibody therapy in breast and colorectal cancer,[12] and more.[13] Additional current and past efforts exist, such as those belonging to the Cancer Biomarkers Research Group of the National Cancer Institute Division of Cancer Prevention, including the Alliance of Glycobiologists for Cancer Research,[14] the Consortium for Imaging and Biomarkers (CIB),[15] and others,[16] and more biomarker information can be obtained directly from the literature or other published sources. Although not always biomarker centric, the National Cancer Institute Informatics Technology for Cancer Research (ITCR) program funds a number of resources, including HemOnc,[17] the Cancer Proteome Atlas (TCPA),[18] the Patient-Specific Drug-Gene Networks for Recommending Targeted Therapies (CDGNet),[19] and others, that describe, generate, analyze, or link cancer data that could provide additional evidence for biomarker relevance. It seems logical, then, that these various data could be combined into a meta-resource for the search and exploration of cancer biomarker–related information.

However, the above-mentioned data sets are highly variable in scope, developed for specific biological applications, specific to certain cancer types, limited by regulatory status, or pertinent to specific clinical populations. Furthermore, data formats are frequently heterogeneous, lacking common attributes to enable integration. Improved modeling of cancer biomarker data would facilitate the integration of interscope biomarker data. Ontology unification and mapping through common resource accessions would improve the quality of integration, downstream assertions, and reasoning. Improved data provenance tracking would transparently communicate parametric configurations and assumptions made during processing and retrieval, and the availability of integrated data (searchable via Web portal designed with key user input) would enhance the usability of the underlying biomarker data.

To address these issues, the OncoMX group developed a data model to integrate public biomarker data from EDRN and the US Food and Drug Administration (FDA) and additional related data around persistent accessions and identifiers. The resulting model provides the foundation for integration of heterogeneous biomarker evidence (using the BioCompute Object [BCO] framework[20,21]) into the OncoMX knowledgebase and Web portal for cancer biomarker exploration. Integrated data types include cancer mutation, cancer differential expression (mRNA and miRNA), cancer expression specificity (single cell RNA [scRNA] sequencing), healthy expression (mRNA from human and mouse), literature mining for cancer mutation and expression, EDRN biomarkers, and FDA-approved breast cancer biomarkers[2].

METHODS

Figure 1 shows an overview of OncoMX integration.

Data Retrieval

Data were provided as .txt, .tsv, or .csv files for the following: literature mining for differential expression in cancer from DEXTER—Disease-Expression Relation Extraction from Text[22] and mutation in cancer from Extraction of Mutation Association to Diseases (DiMeX) (University of Delaware)[23]; RNA sequencing–derived healthy expression calls and ranks for human and mouse from Bgee, a Database for Gene Expression Evolution (SIB Swiss Institute of Bioinformatics)[24]; and the FDA-approved breast cancer biomarkers and cancer cell–type expression specificity from scRNA sequencing (George Washington University). Data were pulled from external consortium databases for public biomarker data from EDRN,[10] cancer mutation data from BioMuta,[25,26] cancer differential expression of mRNA and miRNA from BioXpress,[26,27] and Reactome.[28] The Data Supplement provides a summary of data access and pertinent details for OncoMX and contributing resources.

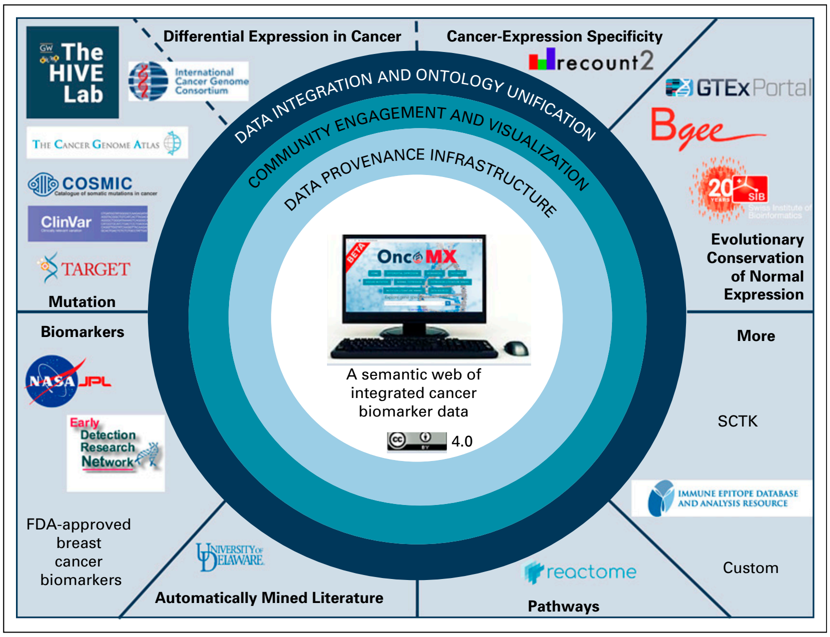

Differential Expression in Cancer
Cancer-Expression Specificity
recount2
GTEx Portal
Bgee
Evolutionary Conservation of Normal Expression
More
SCTK
IMMUNE EPITOPE DATABASE AND ANALYSIS RESOURCE
Custom
reactome
Pathways
Automatically Mined Literature
University of Delaware
FDA-approved breast cancer biomarkers
Early Detection Research Network
NASA JPL
Biomarkers
Mutation
TARGET
ClinVar
COSMIC
The Cancer Genome Atlas
International Cancer Genome Consortium
The HIVE Lab
DATA INTEGRATION AND ONTOLOGY UNIFICATION
COMMUNITY ENGAGEMENT AND VISUALIZATION
DATA PROVENANCE INFRASTRUCTURE
OncoMX
A semantic web of integrated cancer biomarker data
4.0

FIG 1. OncoMX overview. External segments: Data for cancer mutation, cancer expression, healthy expression, literature mining for both cancer mutation and expression, cancer biomarkers, relevant ontologies, functional annotations, and pathways were initially harvested from a number of publicly available resources, such as The Cancer Genome Atlas, International Cancer Genome Consortium, ClinVar, COSMIC, UniProt, and others, and further processed or analyzed by The George Washington University for BioMuta and BioXpress; SIB (Swiss Institute of Bioinformatics) for Bgee (RNA sequencing–derived healthy expression calls and ranks for human and mouse); the University of Delaware for literature mining through DiMeX and DEXTER; and the Early Detection Research Network for biomarkers. Outer ring: Data sets were integrated and unified through Cancer Disease Ontology slim terms for disease names and Uberon Anatomical Entity terms for tissue and physiologic location. Middle ring: Feedback was solicited through a multipronged approach involving use case collection at poster sessions and two formal workshops. With user community guidance, a series of data views and graphic visualizations were devised to aid end users in the exploration and interpretation of biomarker evidence. Inner ring: Data sets were refined and documented with provenance details following the BioCompute Object model for data provenance capture and hosted through the data.oncomx.org data site. Inner circle: The product of these efforts is the current OncoMX Web portal, a semantic web of integrated cancer biomarker data that is readily accessible and licensed under a Creative Commons Attribution 4.0 International License. FDA, US Food and Drug Administration; SCTK, Single-Cell Toolkit.[40]

Processing and Unification

Many data sets used by OncoMX were initially processed by external pipelines documented in source resources. A summary of relevant preprocessing details for cancer mutation, cancer differential expression, healthy expression, and biomarkers is included in the Data Supplement, along with links to documentation. Data sets were unified through Cancer Disease Ontology (CDO) slim[29] terms for disease names and through Uberon Anatomical Entity Terms[30] for anatomic structure names. Data not already annotated with the UniProt Knowledgebase (UniProtKB)[31] accessions (AC) were mapped to the canonical set of UniProtKB AC through gene symbol, transcript, biomarker name, or genes included in a biomarker panel, and AC were filtered if existing outside the canonical set. An OncoMX field dictionary was created that leveraged field names and ontology terms of existing resources (UniProtKB, GlyGen,[32] HUGO Gene Nomenclature Committee (HGNC),[33] and so on), and headers were uniformly formatted and mapped to existing OncoMX field names. Resulting files were checked for completeness and adherence to expected content and subjected to automated quality check to ensure integrity and format sanity. Pipeline steps, provenance details, and other metadata were captured using the BCO[20] specification and output in JavaScript Object Notation (JSON) and .txt formats. All processed data and corresponding metadata objects were entered into a database and subjected to version control. Additional processing considerations for specific data sets are described below.

TABLE 1. Data Sets Available at http://data.oncomx.org

Database ID	Object Description
000003	FDA breast cancer biomarkers
000004	Gene and miRNA expression in cancer from literature (DEXTER)
000007	Cancer nsSNVs (BioMuta version 4.0 full data dump)
000009	Human cancer (CDO slim) to tissue ontology (Uberon) mapping
000012	Genes normally expressed in human tissues (Bgee)
000013	Genes normally expressed in mouse tissues (Bgee)
000015	Cancer cell type expression specificity table
000016	Human cancer driver mutations table
000017	Mapped somatic and germline mutations in cancer
000018	Human cancer mutation mapping (TCGA/ICGC)
000019	EDRN public cancer biomarkers
000020	Mutation in all cancer from literature (DiMeX)
000021	Comprehensive map file for cancer gene symbols, UniProt accessions, RefSeq accessions, taxids
000022	Cancer differentially expressed genes and miRNAs (BioXpress version 4.0 full data dump)
000023	Genes normally expressed in human tissues (Customized)

Abbreviations: CDO, Cancer Disease Ontology; EDRN, Early Detection Research Network; FDA, US Food and Drug Administration; ICGC, International Cancer Genome Consortium; TCGA, The Cancer Genome Atlas.

Mutation and differential expression in cancer. Earlier versions of the cancer mutation and differential expression data sets were updated to include additional data (from The Cancer Genome Atlas, International Cancer Genome Consortium, and Clinical Interpretations of Variants in Cancer [CIViC]), and the expression data set was updated to include miRNA analysis.[26] A new pipeline was developed for cancer mutation to improve the tracking of downloaded content, unify data in .vcf format, subject resulting information to automated quality check, automate mapping of disease terms to CDO slim terms and genes/transcripts to UniProtKB AC using gene symbols or alternate accessions, generate frequency across sources, and automate site annotation and functional impact prediction. Minor updates were made to the differential expression analysis pipeline, including updating the analysis package to DESeq2,[34] mapping cancer subtypes to parent CDO slim terms, and quantifying patient ratios for each direction of expression change (the Data Supplement provides additional details). Resulting data were subjected to the general OncoMX processing steps outlined above before integration.

Healthy expression. A custom format of healthy (wild type) expression data was devised for RNA sequencing–derived expression calls for human and mouse. Data were limited to tissue types that had at least one reported association in the cancer mutation or differential expression data sets, as determined through mapping between CDO slim and Uberon Anatomical Entity. Data were further limited to adult developmental life stages. Analyzed expression values were ranked in two series for each species: one compared the expression of a given gene with that of all genes in a given tissue and the other compared the expression of a given gene with that of the same gene across all tissues. To enable cross-species exploration of expression profiles, the set of 1:1 orthologs between human and mouse were retrieved from the Orthologous MAtrix (OMA) database[35] and used to map human genes to orthologous mouse genes (the Data Supplement provides additional details).

Literature mining. DEXTER,[22] an automated text-mining tool, was applied on Medline abstracts to extract gene, disease, expression level, experimental context, and conditions being compared. Customized for OncoMX, DEXTER extracted expression differences in cancer compared with normal/control tissues and verified whether the normal tissue came from the same patient or not. DEXTER was applied on a comprehensive set of cancer-related abstracts identified using the PubMed query “cancer OR cancers OR carcinoma OR carcinomas OR neoplasm OR neoplasms,” which returned 3,717,745 abstracts (as of March 2018). Abstracts were filtered for those that contained words or phrases pertinent to expression, which reduced the number of abstracts to 1,750,928.

DiMeX[23] is a tool that detects different types of connections between mutations and disease in the literature. For OncoMX, less reliable connections, such as those extracted on the basis of comentions, were dropped while maintaining association relations between mutation and a disease aspect. The mutation detection module of DiMeX was

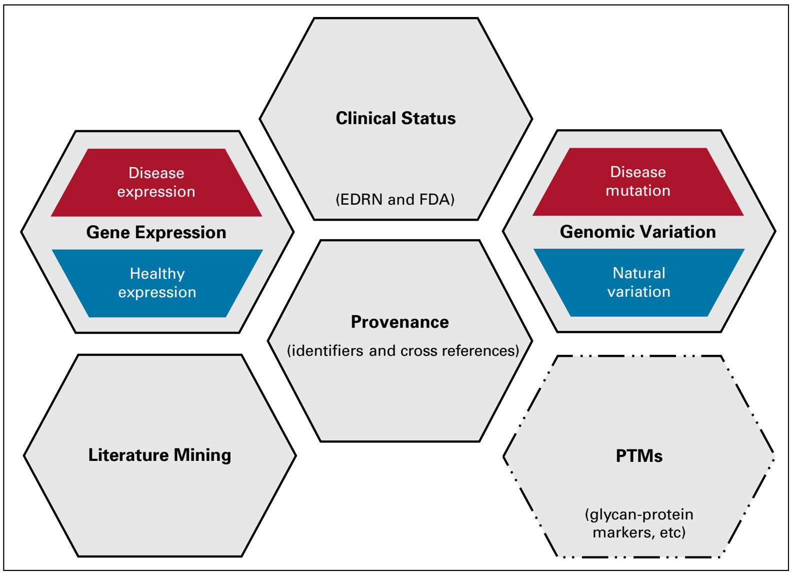
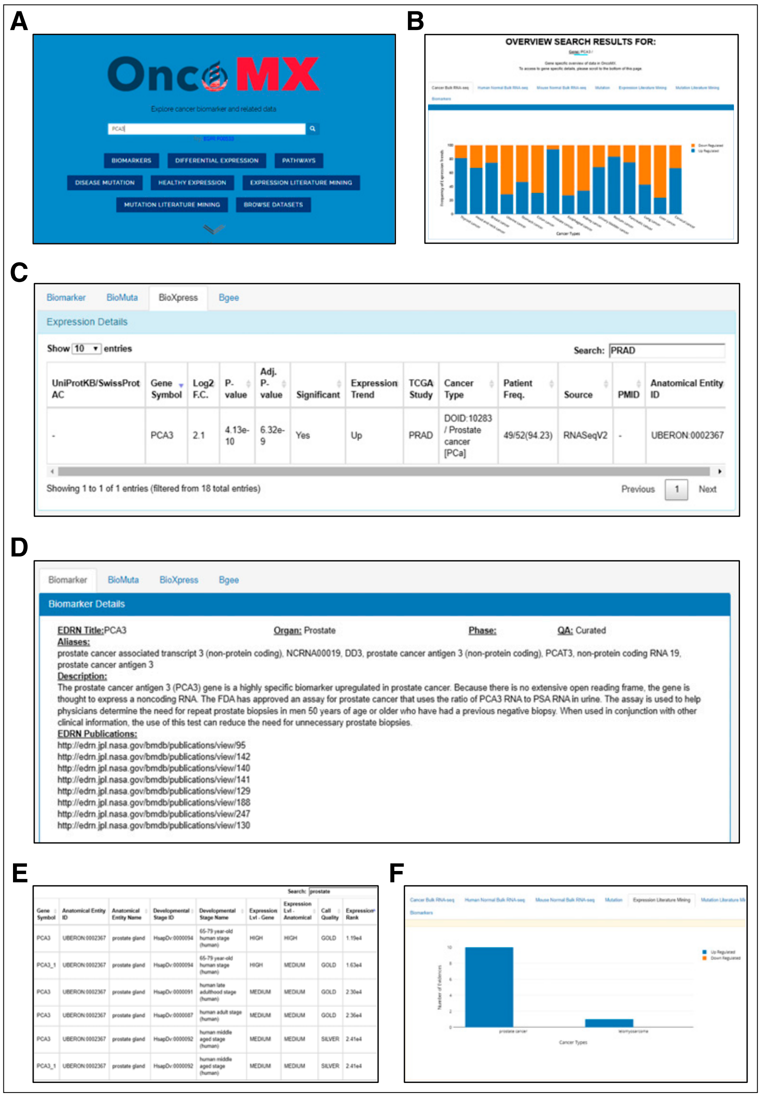


FIG 2. Biomarker evidence data model. This diagram shows the modular extensibility and current configuration of the biomarker evidence data model developed for OncoMX. The central provenance domain component captures biomarker and related identifiers for keys and core attributes, as well as cross references between various dictionaries. The genomic variation component contains both mutations in cancer and natural polymorphism data. The expression component also has two subcomponents to describe data coming from diseased and healthy samples. The healthy subcomponent has an added layer of organism, and both expression subcomponents can be further broken down on the basis of the experimental strategy from which data were generated (not shown). The literature mining component contains data extracted from abstracts and full articles reporting biomarker activity, and the clinical status component contains data from Early Detection Research Network (EDRN) and US Food and Drug Administration (FDA) describing clinical attributes, including approved and actual indications, status of clinical trials, related publications, and more. Of note, extension to a new evidence type, glycan-protein type biomarkers, is actively underway, which should allow for the future extension to other subtypes of the anticipated post-translational modification domain.

improved by integration with tmVar[36] after performance evaluation of several mutation detection tools on a range of well-known corpora. The resulting integration enabled the reduction of false negatives without increasing false positives. DiMeX's relation extraction module was also refined and improved by integrating more recent parsing technology.

Cancer biomarkers. Public biomarker records were retrieved from EDRN, and a new data set of FDA-approved biomarkers in breast cancer was generated by manual search across Web resources (details are provided in the Data Supplement). The generation and integration of these data sets enabled the OncoMX team to troubleshoot issues of extension, facilitated augmentation of infrastructure to ensure interoperability between data of differing types, and demonstrated the immediate utility of newly added data and views available to the end user. A new, easily extensible, biomarker-centered data model was developed to describe a biomarker evidence data object, accounting for the relationships between data sets reporting clinical status of known and/or actively studied biomarkers (EDRN and/or FDA) and other biomarker evidence data types.

Cancer expression specificity. Expression of 10 cancer cell types across the brain, lung, and colon were analyzed from scRNA sequencing data from three studies retrieved from recount2[37] and integrated into OncoMX (details are provided in the Data Supplement). Count tables were filtered for low-quality cells and low-abundance genes by filtering out samples with library sizes and features that expressed fewer than three median absolute deviations from the median value of each quality control metric computed from mitochondrial gene expression across all samples. Cell-specific biases between samples were normalized with a deconvolution approach[38] in which pooled counts from multiple samples are used to compute size factors that are deconvolved to infer size factors for each sample. Data sets were filtered for target cell sequencing runs, and dimensionality reduction was performed using principal component analysis (PCA), t-distributed stochastic neighbor embedding, and uniform manifold approximation and projection (UMAP). scRNA sequencing gene-level count matrices were analyzed for cell type–specific expression using Preferential Expression Measure (PEM), a metric designed for tissue-level expression specificity analysis,[39] and qualitative annotations of specificity were computationally determined for each PEM score. Resulting data were mapped to UniProtKB AC, HGNC gene IDs, Ensembl gene IDs,[41] DOIDs, and DO names, and were subjected to OncoMX processing steps described above.



Expression Details

Show 10 entries

Search: PRAD

UniProtKB/SwissProt AC	Gene Symbol	Log2 F.C.	P-value	Adj. P-value	Significant	Expression Trend	TCGA Study	Cancer Type	Patient Freq.	Source	PMID	Anatomical Entity ID
-	PCA3	2.1	4.13e-10	6.32e-9	Yes	Up	PRAD	DOID:10283 / Prostate cancer [PCa]	49/52(94.23)	RNASeqV2	-	UBERON:0002367

Showing 1 to 1 of 1 entries (filtered from 18 total entries)

Previous 1 Next

Biomarker Details

EDRN Title:PCA3 **Organ:** Prostate **Phase:** **QA:** Curated

Aliases:
prostate cancer associated transcript 3 (non-protein coding), NCRNA00019, DD3, prostate cancer antigen 3 (non-protein coding), PCAT3, non-protein coding RNA 19, prostate cancer antigen 3

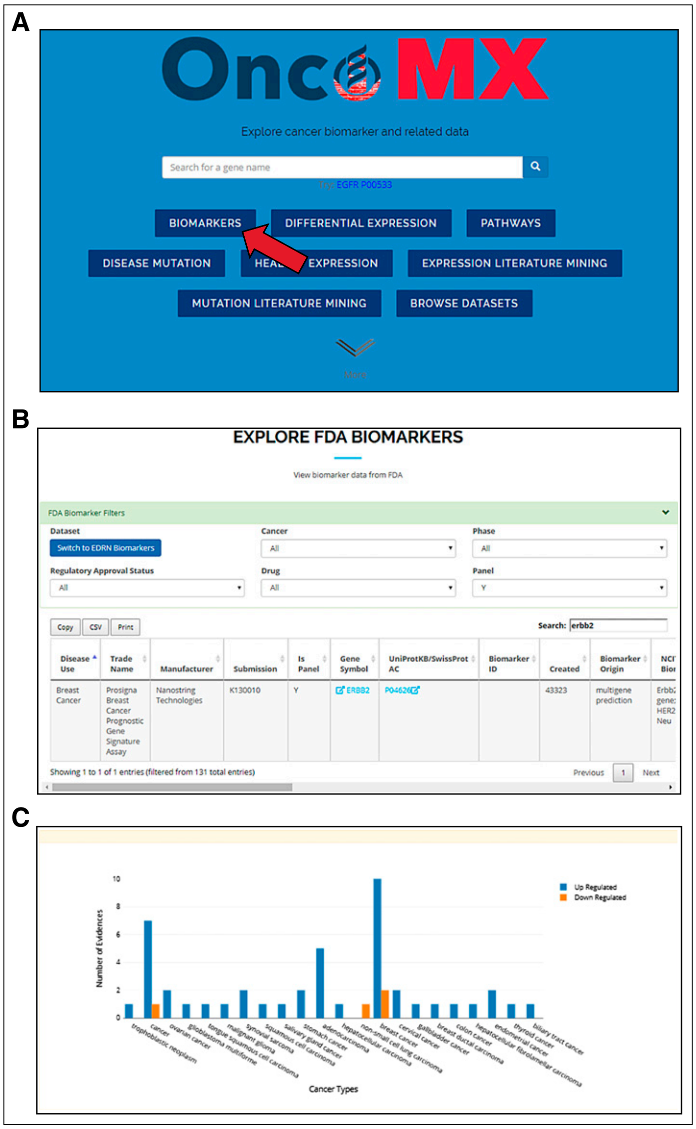
Description:
The prostate cancer antigen 3 (PCA3) gene is a highly specific biomarker upregulated in prostate cancer. Because there is no extensive open reading frame, the gene is thought to express a noncoding RNA. The FDA has approved an assay for prostate cancer that uses the ratio of PCA3 RNA to PSA RNA in urine. The assay is used to help physicians determine the need for repeat prostate biopsies in men 50 years of age or older who have had a previous negative biopsy. When used in conjunction with other clinical information, the use of this test can reduce the need for unnecessary prostate biopsies.

EDRN Publications:
http://edrn.jpl.nasa.gov/bmdb/publications/view/95
http://edrn.jpl.nasa.gov/bmdb/publications/view/142
http://edrn.jpl.nasa.gov/bmdb/publications/view/140
http://edrn.jpl.nasa.gov/bmdb/publications/view/141
http://edrn.jpl.nasa.gov/bmdb/publications/view/129
http://edrn.jpl.nasa.gov/bmdb/publications/view/188
http://edrn.jpl.nasa.gov/bmdb/publications/view/247
http://edrn.jpl.nasa.gov/bmdb/publications/view/130

Search: prostate

Gene Symbol	Anatomical Entity ID	Anatomical Entity Name	Developmental Stage ID	Developmental Stage Name	Expression Lvl - Gene	Expression Lvl - Anatomical	Call Quality	Expression Rank
PCA3	UBERON:0002367	prostate gland	HsapDv:0000094	65-79 year-old human stage (human)	HIGH	HIGH	GOLD	1.19e4
PCA3_1	UBERON:0002367	prostate gland	HsapDv:0000094	65-79 year-old human stage (human)	HIGH	MEDIUM	GOLD	1.63e4
PCA3	UBERON:0002367	prostate gland	HsapDv:0000091	human late adulthood stage (human)	MEDIUM	MEDIUM	GOLD	2.30e4
PCA3	UBERON:0002367	prostate gland	HsapDv:0000087	human adult stage (human)	MEDIUM	MEDIUM	GOLD	2.36e4
PCA3	UBERON:0002367	prostate gland	HsapDv:0000092	human middle aged stage (human)	MEDIUM	MEDIUM	SILVER	2.41e4
PCA3_1	UBERON:0002367	prostate gland	HsapDv:0000092	human middle aged stage (human)	MEDIUM	MEDIUM	SILVER	2.41e4

FIG 3. Exploring OncoMX for an individual gene biomarker: Findings for *PCA3* in prostate cancer. (A) The OncoMX landing page provides quick access to the search bar. Entering "PCA3" will redirect the user to the search results page. (B) The default open tab of the top viewer in the search results page shows mRNA differential expression results for the queried gene across all cancers. *PCA3* is shown to be upregulated in 94% of prostate cancer tumor samples compared with the corresponding adjacent normal samples. (C) Tabular and text details are also available in the lower viewer on the search result pages. Exploring the quantitative information shows that 49 of 52 samples have a logtwo-fold change increase of expression in tumor samples compared with the adjacent normal, and that the overexpression reported for prostate cancer is statistically significant. (D) Text details, including a list of biomarker aliases, a brief description, and a list of links to publications reported by Early Detection Research Network (EDRN), are available from the Biomarkers tab in the lower viewer. (E) Healthy expression of *PCA3* is found to be "MEDIUM" in earlier human adult stages but "HIGH" in the 65- to 79-year-old human stage. (F) Automatic literature mining finds 16 unique sentences from 15 unique publications reporting overexpression of *PCA3* in prostate cancer and another sentence reporting overexpression of *PCA3* in leiomyosarcoma. FDA, US Food and Drug Administration; PSA, prostate-specific antigen; RNA-seq, RNA sequencing; TCGA, The Cancer Genome Atlas.



EXPLORE FDA BIOMARKERS

View biomarker data from FDA

FDA Biomarker Filters

Dataset: Switch to EDRN Biomarkers
Cancer: All
Phase: All
Regulatory Approval Status: All
Drug: All
Panel: Y

Copy CSV Print

Search: erbb2

Disease Use	Trade Name	Manufacturer	Submission	Is Panel	Gene Symbol	UniProtKB/SwissProt AC	Biomarker ID	Created	Biomarker Origin	NCI Bior
Breast Cancer	Prosigna Breast Cancer Prognostic Gene Signature Assay	Nanostring Technologies	K130010	Y	ERBB2	P04626		43323	multigene prediction	Erbb2 gene: HER2 Neu

Showing 1 to 1 of 1 entries (filtered from 131 total entries)

Previous 1 Next

FIG 4. Exploring OncoMX for a gene panel biomarker: Findings for *ERBB2* as part of the Prosigna multigene prediction panel for breast cancer. (A) In addition to the search bar, the OncoMX landing page contains buttons that are quick links to other tables and views. Clicking Biomarkers will redirect the user to the biomarker exploration table viewer. (B) From the data exploration table, the user can click the arrow in the top right of the Biomarker Filters box to access the dropdown menu for filters. Clicking Switch to FDA Biomarkers will reload the table with the US Food and Drug Administration (FDA)–approved biomarker data set. Searching for "ERBB2" in the search bar immediately above the table will reload the table for hits to *ERBB2* (10 hits). Accessing the filters again, the user can search for hits belonging only to panels by selecting Y from the Panel filter dropdown. The table will once again reload to display the single hit for *ERBB2* in a panel, identifying it as part of the Prosigna multigene prediction panel for breast cancer. (C) Going back to the landing page and performing a gene search for "ERBB2" will redirect the user to the detailed search results page. Navigating to the Expression Literature Mining tab, one can readily see the standout peak indicating multiple literature evidences (n = 262) for upregulation of *ERBB2* in breast cancer, including 253 unique sentences from 248 publications. EDRN, Early Detection Research Network.

Integration and Data Modeling

All data were mapped to UniProtKB AC directly or through HGNC gene ID and/or Ensembl gene ID to UniProtKB AC. A data model describing a biomarker evidence object was devised such that an object of type biomarker evidence can be diverse (molecular characteristic, clinical relevance, descriptive literature, and so on) and is currently expected to map to information about one or many genes.

Database Architecture, Back-End Infrastructure, and Front-End Implementation

OncoMX was developed within the virtual machine environment, configured using a 32-GB ram and 12 CPU virtual machine running under the CentOS Linux 7 operating system. The database was built under the MariaDB version 5.5.60 engine, running under the same server as and connected directly to the application, developed using Python Django framework integrating bootstrap 3 for user interface design and jQuery for interactive capabilities. The application was designed with user login security based on the Django authentication library and built to use RESTful Web services to expose the underlying database for external access. The application environment, including the Django framework,[42] was dockerized into a container, which allowed easy transfer and deployment across servers while maintaining a consistent architectural environment.

RESULTS

Integrated Data

Resulting data sets encompassed 939 and 96 unique biomarkers from EDRN and FDA, respectively, mapped to 20,576 genes with available mutation and/or differential expression in cancer. Sentences reporting mutation in cancer were extracted from 14,360 PubMed publications, and those reporting differential expression in cancer were extracted from 25,865 publications. Healthy expression data for 33,753 human and 45,879 mouse Ensembl gene IDs were retrieved, mapping to 19,555 canonical human genes, 15,349 with mouse orthologs. Table 1 lists data sets currently available at https://data.oncomx.org.

Biomarker Evidence Data Modeling

Integration of EDRN and FDA data sets solidified the need to describe both clinical assertions related to actively characterized cancer biomarkers along with other information that could be leveraged as biomarker evidence. A data model was developed to describe a biomarker evidence object that is extensible to any number of components but currently has components for provenance, genomic variation, gene expression, clinical status, and literature mining. Although not required, the current core of the provenance domain is the gene symbol as available biomarker related data contain or link to one or many genes or miRNAs. Granularity of biomarker-level details was retained without sacrificing gene-level detail for panels that contained multiple genes by mapping to each unique gene. The resulting model (Fig 2) accommodates all OncoMX data types and is readily extensible to new data types containing some combination of keys and additional core attributes of disease and tissue; data types not mapped to genes can still be connected through other core attributes.

Usage and Utility

OncoMX Web site features. Early Web development focused on basic gene search functionality such that a string search query results in the display of all visual and tabular information for a specified gene. Combinatorial filtering and analysis on the integrated data layer allow filtering by P value thresholds or specific disease labels. Other features include a landing page with search bar and quick links for data set exploration; a dashboard for visualizing statistics and summaries; a table viewer rendered dynamically on the basis of user interaction; various types of user documentation; and a contact form to solicit user feedback.

Use case 1: Search for an individual gene biomarker. *PCA3* is a long, noncoding RNA that is overexpressed in most prostate cancer PCa cells and is involved in regulating the expression of epithelial–mesenchymal transition markers, androgen receptor cofactors, and PCa suppressor *PRUNE2*.[43-45] As depicted in Figure 3, search for “PCA3” shows an overexpression of *PCA3* in prostate cancer. In fact, 49 of 52 samples have a logtwo-fold change increase in *PCA3* expression in the tumor sample compared with its adjacent normal. Biomarker description, aliases, and publications reported by EDRN can be viewed in the biomarker details tab. Moreover, there are 14 literature evidences reporting upregulated *PCA3* in prostate cancer. Of note, *PCA3* expression in healthy samples is qualified as “MEDIUM” for earlier adult stages, but “HIGH” for later stages, which implies an association between *PCA3* expression and age.

Use case 2: Search for a biomarker panel. The Prosigna Breast Cancer Prognostic Gene Signature Assay panel includes 58 genes that can be used to assess a patient’s risk of disease recurrence (https://www.nanostring.com/diagnostics/prosigna/technology/prosigna-algorithm). Clicking on the Biomarkers button from the landing page will redirect the user to the data exploration table that can be toggled between biomarker data sets and filtered for various attributes. Search for “ERBB2” in this table shows that this gene, well known for its indications for treatment and prognosis in breast cancer,[46-48] is included in 10 FDA-approved biomarker products—nine individual amplification markers and one panel, the Prosigna multigene prediction panel. In the detailed search view, there are multiple evidences for *ERBB2* in the literature, both for differential expression and mutation. Of interest, whereas the majority of differential expression evidences (n = 48) indicate upregulation, there are two evidences for downregulation. Exploration of the underlying data shows that downregulation has been observed in response to pertuzumab treatment in mouse xenograft studies[49] and has been observed to be lower in obese patients with early-stage breast cancer.[50] Visualization of available information shows that this gene is approved for various indications, including prognosis, prediction, and companion diagnosis (Fig 4). The Data Supplement describes additional use cases.

DISCUSSION

Leveraging Existing Resources to Promote Sustainability

OncoMX was designed to combine existing and newly generated data, establishing a strong application customized for cancer biomarker research. Although other multidimensional, integrated cancer resources exist, such as the (cBioPortal)[51,52] and CIViC,[53] the focus on biomarkers, foundational inclusion of large-scale literature mining, ontology-driven unification, and cross comparison of analysis of healthy samples across species are aspects that are unique to OncoMX. However, OncoMX leverages the utility of such extant related resources, including cBioPortal for Cancer Genomics (cBioPortal),[51,52] CDGnet (CANCER Drug Gene Network),[19] CDSA (Cancer Digital Slide Archive),[54] CIViC,[53] HemOnc,[17] iPTMnet,[55] PDX (patient-derived tumor xenograft) finder,[56] and TCPA (The Cancer Proteome Atlas). This model promotes the extensibility and sustainability of OncoMX and referenced resources by allowing OncoMX to focus on harmonization, integration, annotation mapping, evidence tagging, and database maintenance while outsourcing data set creation and curation to collaborators and other experts.

Future Directions

OncoMX is actively seeking new data and types, such as imaging, glycan biomarkers, drug targets, methylation, alternative splicing, and more. Work in progress includes extending the data model to new types, integrating FDA data sets for additional cancers upon user request, improving healthy expression and cross-species visualization, and expanding cross references to key cancer resources. The OncoMX team will continue hosting workshops to ensure that development is shaped by users, facilitating engagement between stakeholders and ultimately resulting in an adaptable resource for the improved exploration and potential discovery of cancer biomarkers.

AFFILIATIONS

[1]The George Washington University, Washington DC
[2]SIB Swiss Institute of Bioinformatics, Lausanne, Switzerland
[3]Department of Ecology and Evolution, University of Lausanne, Lausanne, Switzerland
[4]University of Delaware, Newark, DE
[5]NASA Jet Propulsion Laboratory, Pasadena, CA

CORRESPONDING AUTHOR

Hayley M. Dingerdissen, MS, The George Washington University, 2300 I St NW, Rome 541, Washington DC 20037; e-mail: hmhamilt@gwmail.gwu.edu.

PRIOR PRESENTATION

This work is from a dissertation to be presented to the above program in partial fulfillment of the requirements of a PhD degree (H.M.D.).

SUPPORT

Funded by National Cancer Institute Grant No. U01CA215010 and Early Detection Research Network Associate Member Agreement No. 156620 from the National Cancer Institute.

AUTHOR CONTRIBUTIONS

Conception and design: Hayley M. Dingerdissen, K. Vijay-Shanker, Marc Robinson-Rechavi, Heather Kincaid, Daniel J. Crichton, Raja Mazumder
Administrative support: Amanda Bell
Provision of study materials or patients: Heather Kincaid
Collection and assembly of data: Hayley M. Dingerdissen, Frederic Bastian, K. Vijay-Shanker, Amanda Bell, Nikhita Gogate, Samir Gupta, Evan Holmes, Charles Hadley King, David Liu
Data analysis and interpretation: Hayley Dingerdissen, Frederic Bastian, Amanda Bell, Nikhita Gogate, Samir Gupta, Evan Holmes, Robel Kahsay, Jonathon Keeney, Charles Hadley King, Raja Mazumder
Manuscript writing: All authors
Final approval of manuscript: All authors
Accountable for all aspects of the work: All authors

AUTHORS' DISCLOSURES OF POTENTIAL CONFLICTS OF INTEREST

The following represents disclosure information provided by authors of this manuscript. All relationships are considered compensated unless otherwise noted. Relationships are self-held unless noted. I = Immediate Family Member, Inst = My Institution. Relationships may not relate to the subject matter of this manuscript. For more information about ASCO's conflict of interest policy, please refer to www.asco.org/rwc or ascopubs.org/cci/author-center.

Open Payments is a public database containing information reported by companies about payments made to US-licensed physicians (Open Payments).

Marc Robinson-Rechavi
Employment: Debiopharm Group (I)
Travel, Accommodations, Expenses: Debiopharm Group (I)

Raja Mazumder
Research Funding: Merck, Otsuka

No other potential conflicts of interest were reported.

ACKNOWLEDGMENT

The authors thank Quan Wan and Yu Hu for early contributions to the cancer expression data sets; Julien Wolbrett for early contributions to healthy expression data sets; Ashique Mahmood for early contributions to literature mining data sets; Lynn Schriml for guidance and expertise with respect to ontology mapping; Lama ElZohary for additional work on ontology mapping; Reza Mousavi Jafarabad for early assistance preparing scripts for expression analysis; Brian Fochtman for scripts and expression analysis; Dara Baker and Manna Kidane for creation and curation of US

Food and Drug Administration biomarker data sets; Stephanie Singleton for early drafts of user documentation and critical administrative assistance; Sneh Talwar for assistance preparing draft figures and data tables; Peng Su for contributions to and support for mutation literature mining; Edmund Cauley and Daniel Lyman for manuscript review; and Jeet Kiran Vora and Rahi Navelkar for thoughtful discussion regarding data set maintenance and related considerations.

REFERENCES

1. OncoMX: https://www.oncomx.org/
2. National Cancer Institute: NCI dictionary of cancer terms. https://www.cancer.gov/publications/dictionaries/cancer-terms
3. Villalobos P, Wistuba II: Lung cancer biomarkers. Hematol Oncol Clin North Am 31:13-29, 2017
4. Henry NL, Hayes DF: Cancer biomarkers. Mol Oncol 6:140-146, 2012
5. O'Connor JP, Aboagye EO, Adams JE, et al: Imaging biomarker roadmap for cancer studies. Nat Rev Clin Oncol 14:169-186, 2017
6. Montalban-Arques A, Scharl M: Intestinal microbiota and colorectal carcinoma: Implications for pathogenesis, diagnosis, and therapy. EBioMedicine 48:648-655, 2019
7. Goossens N, Nakagawa S, Sun X, et al: Cancer biomarker discovery and validation. Transl Cancer Res 4:256-269, 2015
8. Kawazoe A, Shitara K: Next-generation sequencing and biomarkers for gastric cancer: What is the future? Ther Adv Med Oncol 11:1-17, 2019
9. Collins FS, Varmus H: A new initiative on precision medicine. N Engl J Med 372:793-795, 2015
10. National Cancer Institute: Early Detection Research Network. https://prevention.cancer.gov/major-programs/early-detection-research
11. Zhang X, Sun XF, Cao Y, et al: CBD: A biomarker database for colorectal cancer. Database (Oxford) 2018:bay046, 2018
12. Pérez-Granado J, Piñero J, Furlong LI: ResMarkerDB: A database of biomarkers of response to antibody therapy in breast and colorectal cancer. Database (Oxford) 2019:baz060, 2019
13. Dienstmann R, Jang IS, Bot B, et al: Database of genomic biomarkers for cancer drugs and clinical targetability in solid tumors. Cancer Discov 5:118-123, 2015
14. National Cancer Institute: Alliance of Glycobiologists for Cancer Research. https://prevention.cancer.gov/major-programs/alliance-glycobiologists
15. National Cancer Institute: Consortium for Imaging and Biomarkers. https://prevention.cancer.gov/major-programs/consortium-imaging-and
16. National Cancer Institute: Cancer biomarkers. https://prevention.cancer.gov/research-groups/cancer-biomarkers
17. Warner JL, Dymshyts D, Reich CG, et al: HemOnc: A new standard vocabulary for chemotherapy regimen representation in the OMOP common data model. J Biomed Inform 96:103239, 2019
18. Li J, Akbani R, Zhao W, et al: Explore, visualize, and analyze functional cancer proteomic data using the cancer proteome atlas. Cancer Res 77:e51-e54, 2017
19. Kancherla J, Rao S, Bhuvaneshwar K, et al: Evidence-based network approach to recommending targeted cancer therapies. JCO Clin Cancer Infor , 2020
20. Alterovitz G, Dean D, Goble C, et al: Enabling precision medicine via standard communication of HTS provenance, analysis, and results. PLoS Biol 16: e3000099, 2018
21. Simonyan V, Goecks J, Mazumder R: Biocompute objects: A step towards evaluation and validation of biomedical scientific computations. PDA J Pharm Sci Technol 71:136-146, 2017
22. Gupta S, Dingerdissen H, Ross KE, et al: DEXTER: Disease-Expression Relation Extraction from Text. Database (Oxford) 2018:bay045, 2018
23. Mahmood AS, Wu TJ, Mazumder R, et al: DiMeX: A text mining system for mutation-disease association extraction. PLoS One 11:e0152725, 2016
24. Bastian F, Parmentier G, Roux J, et al. Bgee: Integrating and comparing heterogeneous transcriptome data among species, in Bairoch A, Cohen-Boulakia S, Froidevaux C (eds): Data Integration in the Life Sciences. Berlin, Germany, Springer-Verlag, 2008, pp 124-131
25. Wu TJ, Shamsaddini A, Pan Y, et al: A framework for organizing cancer-related variations from existing databases, publications and NGS data using a High-performance Integrated Virtual Environment (HIVE). Database (Oxford) 2014:bau022, 2014
26. Dingerdissen HM, Torcivia-Rodriguez J, Hu Y, et al: BioMuta and BioXpress: Mutation and expression knowledgebases for cancer biomarker discovery. Nucleic Acids Res 46:D1128-D1136, 2018
27. Wan Q, Dingerdissen H, Fan Y, et al: BioXpress: An integrated RNA-seq-derived gene expression database for pan-cancer analysis. Database (Oxford) 2015: bav019, 2015
28. Fabregat A, Jupe S, Matthews L, et al: The reactome pathway knowledgebase. Nucleic Acids Res 46(D1):D649-D655, 2018
29. Wu TJ, Schriml LM, Chen QR, et al: Generating a focused view of disease ontology cancer terms for pan-cancer data integration and analysis. Database (Oxford) 2015:bav032, 2015
30. Mungall CJ, Torniai C, Gkoutos GV, et al: Uberon, an integrative multi-species anatomy ontology. Genome Biol 13:R5, 2012
31. UniProt Consortium: UniProt: A worldwide hub of protein knowledge. Nucleic Acids Res 47:D506-D515, 2019
32. York WS, Mazumder R, Ranzinger R, et al: GlyGen: Computational and informatics resources for glycoscience. Glycobiology 30:72-73, 2020
33. Yates B, Braschi B, Gray KA, et al: Genenames.org: The HGNC and VGNC resources in 2017. Nucleic Acids Res 45:D619-D625, 2017
34. Love MI, Huber W, Anders S: Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 15:550, 2014
35. Altenhoff AM, Glover NM, Train CM, et al: The OMA orthology database in 2018: Retrieving evolutionary relationships among all domains of life through richer web and programmatic interfaces. Nucleic Acids Res 46:D477-D485, 2018
36. Wei CH, Phan L, Feltz J, et al: tmVar 2.0: Integrating genomic variant information from literature with dbSNP and ClinVar for precision medicine. Bioinformatics 34:80-87, 2018
37. Collado-Torres L, Nellore A, Kammers K, et al: Reproducible RNA-seq analysis using recount2. Nat Biotechnol 35:319-321, 2017
38. Lun AT, Bach K, Marioni JC: Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. Genome Biol 17:75, 2016
39. Huminiecki L, Lloyd AT, Wolfe KH: Congruence of tissue expression profiles from Gene Expression Atlas, SAGEmap and TissueInfo databases. BMC Genomics 4:31, 2003
40. Jenkins DF, Faits T, Briars E, et al: Interactive single cell RNA-Seq analysis with the Single Cell Toolkit (SCTK). bioRxiv doi:10.1101/329755 [epub ahead of print on May 24, 2018]
41. Zerbino DR, Achuthan P, Akanni W, et al: Ensembl 2018. Nucleic Acids Res 46:D754-D761, 2018
42. Django. https://www.djangoproject.com/download/

43. Lemos AE, Ferreira LB, Batoreu NM, et al: PCA3 long noncoding RNA modulates the expression of key cancer-related genes in LNCaP prostate cancer cells. Tumour Biol 37:11339-11348, 2016

44. Salameh A, Lee AK, Cardó-Vila M, et al: PRUNE2 is a human prostate cancer suppressor regulated by the intronic long noncoding RNA PCA3. Proc Natl Acad Sci USA 112:8403-8408, 2015

45. Zhang G, He X, Ren C, et al: Long noncoding RNA PCA3 regulates prostate cancer through sponging miR-218-5p and modulating high mobility group box 1. J Cell Physiol 234:13097-13109, 2019

46. Waks AG, Winer EP: Breast cancer treatment: A review. JAMA 321:288-300, 2019

47. Pondé N, Brandão M, El-Hachem G, et al: Treatment of advanced HER2-positive breast cancer: 2018 and beyond. Cancer Treat Rev 67:10-20, 2018

48. D'Amato V, Raimondo L, Formisano L, et al: Mechanisms of lapatinib resistance in HER2-driven breast cancer. Cancer Treat Rev 41:877-883, 2015

49. Lam K, Chan C, Reilly RM: Development and preclinical studies of ^{64}Cu-NOTA-pertuzumab F(ab')$_2$ for imaging changes in tumor HER2 expression associated with response to trastuzumab by PET/CT. MAbs 9:154-164, 2017

50. Mousa U, Onur H, Utkan G: Is obesity always a risk factor for all breast cancer patients? c-erbB2 expression is significantly lower in obese patients with early stage breast cancer. Clin Transl Oncol 14:923-930, 2012

51. Cerami E, Gao J, Dogrusoz U, et al: The cBio cancer genomics portal: An open platform for exploring multidimensional cancer genomics data. Cancer Discov 2:401-404, 2012

52. Gao J, Aksoy BA, Dogrusoz U, et al: Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Sci Signal 6:pl1, 2013

53. Griffith M, Spies NC, Krysiak K, et al: CIViC is a community knowledgebase for expert crowdsourcing the clinical interpretation of variants in cancer. Nat Genet 49:170-174, 2017

54. Gutman DA, Khalilia M, Lee S, et al: The digital slide archive: A software platform for management, integration, and analysis of histology for cancer research. Cancer Res 77:e75-e78, 2017

55. Huang H, Arighi CN, Ross KE, et al: iPTMnet: An integrated resource for protein post-translational modification network discovery. Nucleic Acids Res 46:D542-D550, 2018

56. Conte N, Mason JC, Halmagyi C, et al: PDX Finder: A portal for patient-derived tumor xenograft model discovery. Nucleic Acids Res 47:D1073-D1079, 2019