2017

BioMuta and BioXpress: mutation and expression knowledgebases for cancer biomarker discovery

Hayley Dingerdissen
George Washington University

John Torcivia-Rodriguez
George Washington University

Yu Hu
George Washington University

Ting-Chia Chang
George Washington University

Raja Mazumder
George Washington University

See next page for additional authors

APA Citation

Dingerdissen, H., Torcivia-Rodriguez, J., Hu, Y., Chang, T., Mazumder, R., & Kashay, R. (2017). BioMuta and BioXpress: mutation and expression knowledgebases for cancer biomarker discovery. *Nucleic Acids Research,* (). http://dx.doi.org/10.1093/nar/gkx907

Authors

Hayley Dingerdissen, John Torcivia-Rodriguez, Yu Hu, Ting-Chia Chang, Raja Mazumder, and Robel Kashay

doi: 10.1093/nar/gkx907

BioMuta and BioXpress: mutation and expression knowledgebases for cancer biomarker discovery

Hayley M. Dingerdissen[1,†], John Torcivia-Rodriguez[1,†], Yu Hu[1], Ting-Chia Chang[1], Raja Mazumder[1,2] and Robel Kahsay[1,*]

[1]The Department of Biochemistry & Molecular Medicine, The George Washington University Medical Center, Washington, DC 20037, USA and [2]McCormick Genomic and Proteomic Center, The George Washington University, Washington, DC 20037, USA

Received August 15, 2017; Revised September 21, 2017; Editorial Decision September 25, 2017; Accepted September 26, 2017

ABSTRACT

Single-nucleotide variation and gene expression of disease samples represent important resources for biomarker discovery. Many databases have been built to host and make available such data to the community, but these databases are frequently limited in scope and/or content. BioMuta, a database of cancer-associated single-nucleotide variations, and BioXpress, a database of cancer-associated differentially expressed genes and microRNAs, differ from other disease-associated variation and expression databases primarily through the aggregation of data across many studies into a single source with a unified representation and annotation of functional attributes. Early versions of these resources were initiated by pilot funding for specific research applications, but newly awarded funds have enabled hardening of these databases to production-level quality and will allow for sustained development of these resources for the next few years. Because both resources were developed using a similar methodology of integration, curation, unification, and annotation, we present BioMuta and BioXpress as allied databases that will facilitate a more comprehensive view of gene associations in cancer. BioMuta and BioXpress are hosted on the High-performance Integrated Virtual Environment (HIVE) server at the George Washington University at https://hive.biochemistry.gwu.edu/biomuta and https://hive.biochemistry.gwu.edu/bioxpress, respectively.

INTRODUCTION

Single-nucleotide variations (SNVs) are sequence alterations in DNA that exist between individuals or can accumulate over time within an individual. These variations can be associated with diseases or other phenotypes and hold tremendous value for researchers aiming to characterize the role of variation in disease (1,2). These changes are identified by DNA sequencing in various types of studies, such as genome-wide association studies (GWAS). Despite limitations like variable penetrance and indirect association inherent to the GWAS approach (3), diseases have been successfully linked to multiple, seemingly independent SNVs and such studies have yielded increasingly confident associations with the continued evolution of statistical models (4,5). While nonsynonymous SNVs (nsSNVs), those SNVs that result in altered amino acid sequences, can directly change protein structure and therefore function, gene and microRNA (miRNA) dysregulation can alter normal expression and can also contribute to disease (6,7). Differential expression analysis of RNA-seq data can quantify the expression levels of genes or miRNAs across multiple samples and multiple conditions to identify important markers in disease diagnosis, progression, and treatment (8,9). Because of the potential wealth of clinically relevant information to be gleaned from these data, substantial efforts and resources have been dedicated to host, maintain, and make available to various research communities both normal and suspected disease-associated variation and expression data (2,10–13).

There are many extant databases containing some form of disease variation or expression data, including but not limited to the Database of Single Nucleotide Polymorphisms (dbSNP) (10), the Human Gene Mutation Database (HGMD) (11), ClinVar (14), the Cancer Genome Atlas (TCGA) (https://cancergenome.nih.gov/), CIViC (15), Expression Atlas (16) and others (17–24). However, these current databases are frequently not comprehensive, but rather cover individual studies or a limited number of data adhering to some very strict criterion. Primary repositories like TCGA and the International Cancer Genome Consortium (ICGC) (25) contain raw DNA- and RNA-seq data

[*]To whom correspondence should be addressed. Tel: +1 302 540 8003; Fax: +1 202 994 8974; Email: rykahsay@gwu.edu

[†]These authors contributed equally to this work as first authors.

as well as processed variant calls and expression levels, but rarely contain value-added information regarding the interpretation of individual samples or the study as a whole. Other database projects like cBioPortal (20,21) and CIViC do collect data from numerous studies, but the variety of available data in each is entirely dependent on the intended scope of each project. For example, while CIViC's approach to curation of clinically relevant cancer variants and dedication to transparency ensure the highest quality of included variants, the workflow of manual curation requiring agreement by two independent curators creates an inherent tradeoff between quality and database growth rate potential (15). Our databases, BioMuta (26,27) and BioXpress (28), differ from previous attempts at cataloging disease-associated mutation and expression primarily through the aggregation of data across many studies into a single source with a unified representation. BioMuta is a database of single-nucleotide variations (SNVs) identified in tumor tissue while BioXpress is a database of genes that are differentially expressed in adjacent normal and tumor tissue from the same patient. These databases can be used throughout the research lifecycle, allowing a quick survey of markers previously reported in major studies, by driving hypothesis generation, or by forming the basis for experiments that require more raw data than is available in a single study to support a finding. For example, the BioMuta dataset could be used to discover the most common variations for each cancer, which could then be analyzed to determine the genetic similarity between different cancer types (29).

All knowledgebases generated by our group heavily emphasize the importance of a unified approach to integration, curation, and representation by a single vocabulary (in this case, the subset of cancer disease ontology terms (30)). BioMuta and BioXpress were built following this model by retrieving data primarily from large genomic, public studies, filtering and performing quality control (QC) measures on the data in accordance with our goals, annotating the remaining data with predictions and a diversity of functional annotations, and repackaging the enhanced information into easily usable tables and graphical user interfaces. Early versions of these resources were predominantly datasets used for research by our group and collaborators, sustained through pilot funding. However, strong interest from our user community, demonstrated through high data access traffic and critical feedback, has encouraged us to increase efforts to develop these resources further into production level modules, and to continuously update and increase their content with datasets from new studies.

Although mutation and expression of the same gene are not necessarily mechanistically related, identification of one or both features in disease can strengthen the implication that a given gene or miRNA marker or the pathway(s) involving that marker is likely to be important in that disease. Furthermore, our interaction with users has demonstrated that distinct users have very different research endpoints, many of which are as interested in global implications as in specific gene or miRNA-centric findings. For this reason, we consider the strength of BioMuta and BioXpress in their position as allied databases toward a comprehensive view of cancer involvement, and suggest that each should not be thought of in isolation. We have also implemented fully documented APIs for accessing content on both databases to allow programmatic integration with other resources.

MATERIALS AND METHODS

BioMuta

The BioMuta dataset was generated using a pipeline developed to curate and validate variation data from multiple sources (Figure 1) (26).

Data retrieval, integration, and ID mapping. Raw data was downloaded from several repositories including TCGA, the Catalog of Somatic Mutations in Cancer (COSMIC) (31), ICGC, IntOGen (32) and ClinVar. UnitProtKB/Swiss-Prot (33) resource was used as a source for both variation and annotation data. Some of the data sources offer alternate versions of the same data depending on individual project goals and requirements. For each source, we retrieved the most expansive coding region dataset. For example, while TCGA has conducted both whole genome and exome sequencing, we used only the exome-specific variants for this version of BioMuta (see Supplemental Table S1 for access dates and versions of primary resources). In addition to data sources mentioned, we include a small but important dataset of cancer-associated variations manually gathered from literature review. For each dataset considered, genomic positions were verified using custom QC scripts that validate the reported reference nucleotide at a given genomic position making sure the same nucleotide exists in that position of the reference human genome assembly (GRCh37/hg19). Next, each dataset was reformatted to match the input requirements for the ANNOVAR (34) annotation software tool. The ANNOVAR software tool was used to annotate variants, which are reported in genomic coordinates, with associated gene and protein IDs, strand orientation, position on transcript and protein sequences, and corresponding reference and variant amino acid residues based on annotation in the RefSeq (35,36) database. For increased accuracy, QC measures were performed on ANNOVAR results to validate the reference nucleotide and amino acid at mutation positions matches the nucleotide and amino acid on respective RefSeq sequences.

Up to this point, datasets from each source were analyzed and maintained separately. At this stage, independent datasets were merged and a new record attribute was added to store data source information. All variants in the merged dataset were then mapped to the corresponding reviewed UniProtKB sequences using custom software that uses RefSeq-UniProtKB ID mapping (37) and performs protein sequence alignment using Clustal Omega (38) to validate the reported mapping between the RefSeq sequence and the canonical UniProt isoform and to convert all positions to UniProt coordinates. Another QC procedure was applied here to check if the reported amino acids with UniProt coordinates match the ones at the same postions in UniProt canonical protein sequences.

Functional prediction and annotation. To increase the functional scope of the database, variants were then assessed by PolyPhen (39), a software that predicts the impact on both the structure and function of a protein based on the change

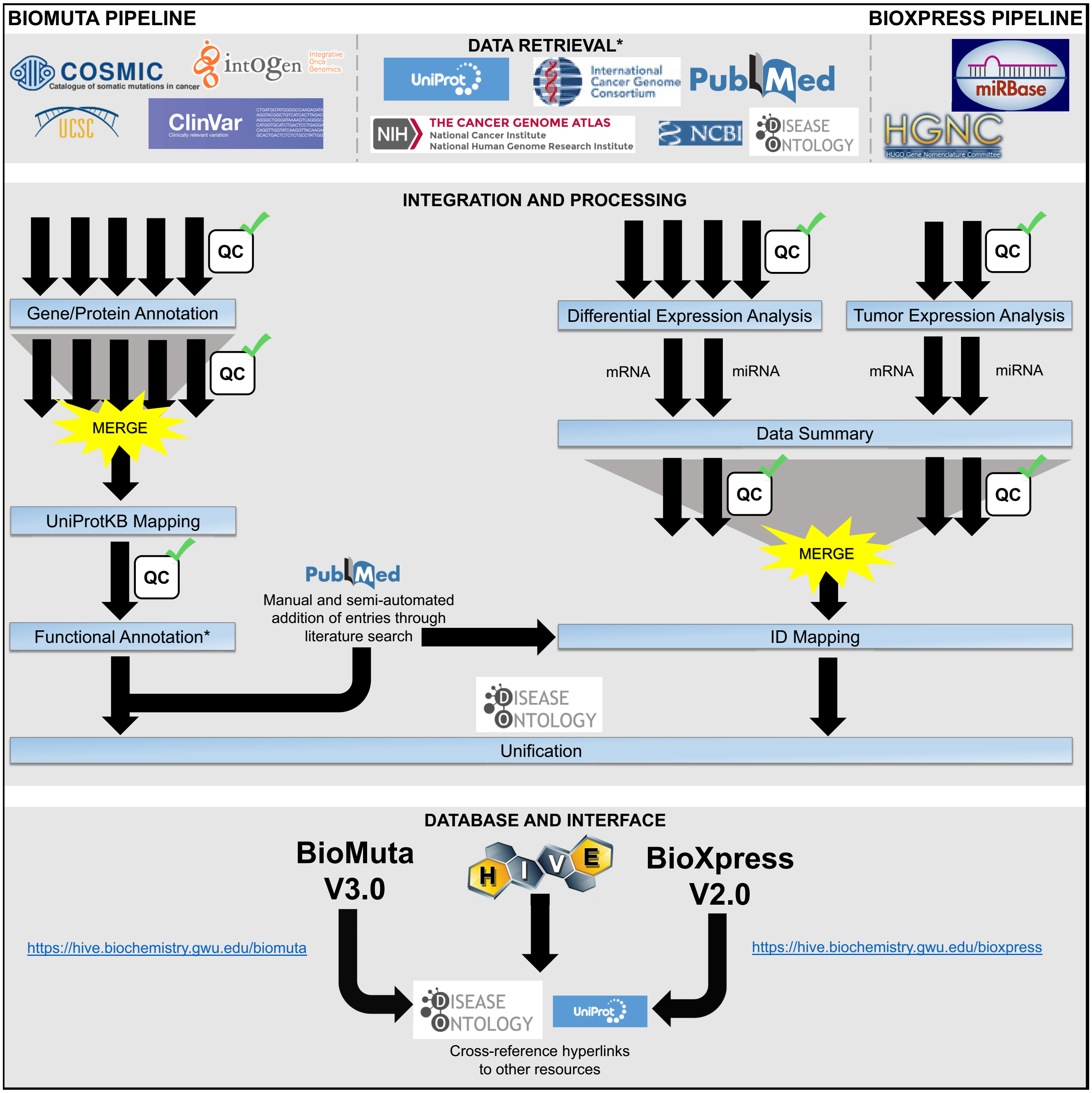

BIOMUTA PIPELINE
BIOXPRESS PIPELINE
DATA RETRIEVAL*
COSMIC Catalogue of somatic mutations in cancer
intOGen Integrative Onco Genomics
UCSC
ClinVar Clinically relevant variation
UniProt
International Cancer Genome Consortium
PubMed
NIH THE CANCER GENOME ATLAS National Cancer Institute National Human Genome Research Institute
NCBI
DISEASE ONTOLOGY
miRBase
HGNC HUGO Gene Nomenclature Committee
INTEGRATION AND PROCESSING
QC
Gene/Protein Annotation
QC
MERGE
UniProtKB Mapping
QC
Functional Annotation*
PubMed
Manual and semi-automated addition of entries through literature search
QC
Differential Expression Analysis
QC
Tumor Expression Analysis
mRNA
miRNA
mRNA
miRNA
Data Summary
QC
QC
MERGE
ID Mapping
DISEASE ONTOLOGY
Unification
DATABASE AND INTERFACE
BioMuta V3.0
HIVE
BioXpress V2.0
https://hive.biochemistry.gwu.edu/biomuta
https://hive.biochemistry.gwu.edu/bioxpress
DISEASE ONTOLOGY
UniProt
Cross-reference hyperlinks to other resources

Figure 1. The development pipelines of both BioMuta and BioXpress share several common features, including primary data sources, integration and ID mapping approaches, unification and interface design. Several sources supply primary data including variation, expression, annotation and ontology/identifier data. In the 'Data Retrieval' portion of the figure, the sources to the far left represent those data sources used only for BioMuta. Similarly, the sources to the far right are those used only for BioXpress. The sources in the middle (between the dashed gray lines) are datasets or sources that contribute data to both BioMuta and BioXpress. Throughout data processing, a number of quality control (QC) steps are imposed to ensure integrity and accuracy of data, where possible. Processed data are unified by cancer type to the corresponding DOID(s) and entered into MySQL database to be searchable by query on the web interfaces. * Due to the number of primary data sources, those resources supplying only functional annotations are not included in the figure above. Sources for functional annotations not pictured include: CDD, SysPTM, PhosphoSite, Phospho.ELM, dbSNO, HPRD and OGlycBase6.0. Additional annotations are supplied following analysis by Polyphen and NetNGlyc.

in amino acid (if any) that occurs during nucleotide substitution. The PolyPhen tool uses a Naïve Bayes approach to determine whether a mutation is benign, possibly damaging, or probably damaging based on several different structural calculations. Similarly, NetNGlyc was used to annotate post-translational modification (PTM) sites at any positions where there are SNVs that affect N-linked glycosylation (primarily based on existence of the consensus NXS/T sequon for N-glycosylation). Note that in its native configuration, NetNGlyc restricts protein length to 4000, preventing annotation of variation sites in proteins of length greater than 4000. Specific annotations for PTMs including glycosylation, phosphorylation and other functional implications (active and binding site) were obtained through retrieval and mapping of UniProtKB feature (FT) lines (see Supplemental Table S2 for a list of all FT entities used in functional annotations). Additional sources of functional annotations include: CDD (40), SysPTM 1.1 (41), PhosphoSite (42), Phospho.ELM (43), dbSNO 1.0 (44), HPRD 9.0 (45) and OGlycBase6.0 (45).

Finally, all variation entries were unified to Disease Ontology or DO (46) terms to facilitate better cancer classification and easier database searching. Specifically, entries were mapped to the subset of DO Cancer Slim (30) terms that represent a streamlined vocabulary for cancer nomenclature.

BioXpress

The BioXpress dataset was constructed through a similar pipeline involving data integration, annotation, unification, analysis, and databasing (Figure 1). For mRNA, raw RNA-seq read counts were downloaded from TCGA by TCGA Assembler (47) restricting the value for assayPlatform to 'gene_RNAseq.' Similarly, for miRNA, the values for assayPlatform were restricted to 'mir_GA.hg19.mirbase20' and 'mir_HiSeq.hg19.mirbase20' (see Supplemental Table S3 for access dates and versions of primary resources).

Differential expression using paired data. Downloaded dataset was filtered for those samples with matched tumor and adjacent non-tumor tissue. The dataset was then used to build an appropriate schema for differential expression analysis using DESeq2 (48). Rows, corresponding to specific transcripts with no raw counts (no expression) in any sample, were deleted prior to differential expression analysis. For each sample for a given cancer type, a model matrix was built with two catgory designations: one category described the disease status for each sample with possible values 'cancer' and 'non-tumor,' the other containing the TCGA patient Ids for the sample. The program DESeq2 then normalized the input data and analyzed differential expression based on the schema labels. From the results of differential expression analysis, over- and under-expression of a transcript in one cancer type were defined as $|\log_2 \text{fold change}| > 0$. For miRNA, the cutoff for significance is an adjusted P-value $< 0.05/n$ (Bonferroni's Approach, where n is the total number of expressed miRNAs in each cancer type). Patient frequencies were also generated including counts of patients with over- or under-expression for each gene/miRNA along with the total number of patients in a given cancer.

All tumor expression using tumor-only data. In addition to the data from TCGA described above, two datasets were retrieved for miRNA from ICGC by the following criteria: select all projects not from the US (most are from TCGA, and the remaining four from TARGET contain only BAM files for miRNAseq data); select data type as 'miRNA-seq'. Normalized read counts for each gene/miRNA from each cancer type were extracted and five quantile values were calculated for plotting boxplots. Patient frequencies and percentiles of tumor expression were calculated by customized R scripts.

ID/cancer-type mapping. For mRNA, RefSeq IDs were extracted from the NCBI-maintained ID mapping (ftp://ftp.ncbi.nlm.nih.gov/refseq/H_sapiens/mRNA_Prot/), and the list of UniProtKB reviewed entries for human was downloaded from UniProtKB resource (ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/idmapping/by_organism/). miRBase IDs for miRNA were also downloaded from the miRBase resource (http://www.mirbase.org/ftp.shtml), and HGNC symbols and Ensembl ID list were retrieved from HGNC site (http://www.genenames.org/cgi-bin/statistics). Annotation and differential expression analysis results from DESeq2 for all datasets were loaded into a relational database to facilitate interpretation of significant dysregulation of genes in cancer, and cancer types were unified by Disease Ontology DOIDs as described for BioMuta.

RESULTS

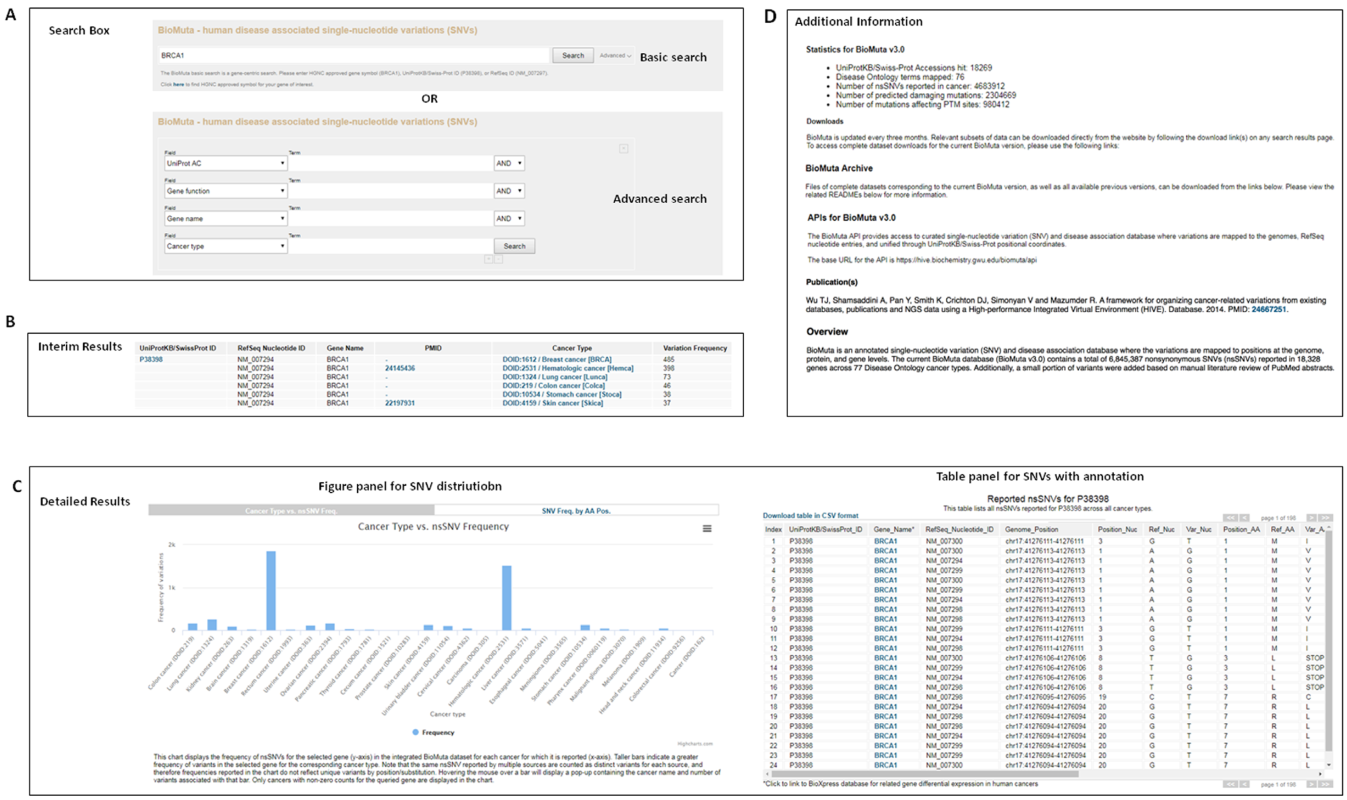
Data summary

Out of 7 373 923 total variants pooled from eight sources, 4 684 236 pass all QC and filtration steps to make it into the final BioMuta database. These variants occur in 18 269 genes and encompass 2599 different cancer terms across all resources, which are then mapped to 77 DO Cancer Slim terms. We find 2 304 757 entries predicted to be damaging by PolyPhen, and 980 447 occurring at a PTM site. 326 (Manual curation) and 1098 (Literature mining) additional entries were added through manual literature review.

Similarly, of 18 596 UniProtKB/Swiss-Prot accessions identified to be expressed in cancer samples, 17 537 genes were reported to be differentially expressed in cancer, which were deemed significant with an associated $P < 0.05$ (corrected for multiple testing) for at least one cancer type. 710 miRNAs were differentially expressed, which were significant.

Interface

Both BioMuta and BioXpress have web-based interfaces (Figures 2 and 3), although the underlying data are available for direct download and accessible through APIs which are fully documented. Web interfaces allow users to directly query the data and represent results in visual form. Both resources offer basic and advanced search options, and also enable users to link internally between resources and externally (i.e. PubMed) as available.


A
Search Box
Basic search
OR
Advanced search
B
Interim Results
C
Detailed Results
Figure panel for SNV distriutiobn
Table panel for SNVs with annotation
D
Additional Information
Search
Advanced
Field
Term
UniProt AC
Gene function
Gene name
Cancer type
AND
Cancer Type vs. nsSNV Freq.
SNV Freq. by AA Pos.
Cancer Type vs. nsSNV Frequency
2k
1k
0
Frequency of variations
Colon cancer (DOID:219)
Lung cancer (DOID:1324)
Kidney cancer (DOID:263)
Brain cancer (DOID:1319)
Breast cancer (DOID:1612)
Rectum cancer (DOID:1993)
Uterine cancer (DOID:363)
Ovarian cancer (DOID:2394)
Pancreatic cancer (DOID:1793)
Thyroid cancer (DOID:1781)
Cecum cancer (DOID:1521)
Prostate cancer (DOID:10283)
Skin cancer (DOID:4159)
Urinary bladder cancer (DOID:11054)
Cervical cancer (DOID:4362)
Carcinoma (DOID:305)
Hematologic cancer (DOID:2531)
Liver cancer (DOID:3571)
Esophageal cancer (DOID:5041)
Meningioma (DOID:3565)
Stomach cancer (DOID:10534)
Pharynx cancer (DOID:0060119)
Malignant glioma (DOID:3070)
Melanoma (DOID:1909)
Head and neck cancer (DOID:11934)
Colorectal cancer (DOID:9256)
Cancer (DOID:162)
Cancer type
Frequency
Highcharts.com
Download table in CSV format
<< < page 1 of 198 > >>

BioMuta - human disease associated single-nucleotide variations (SNVs)

BRCA1

The BioMuta basic search is a gene-centric search. Please enter HGNC approved gene symbol (BRCA1), UniProtKB/Swiss-Prot ID (P38398), or RefSeq ID (NM_007297).

Click here to find HGNC approved symbol for your gene of interest.

BioMuta - human disease associated single-nucleotide variations (SNVs)

UniProtKB/SwissProt ID	RefSeq Nucleotide ID	Gene Name	PMID	Cancer Type	Variation Frequency
P38398	NM_007294	BRCA1	-	DOID:1612 / Breast cancer [BRCA]	485
	NM_007294	BRCA1	24145436	DOID:2531 / Hematologic cancer [Hemca]	398
	NM_007294	BRCA1	-	DOID:1324 / Lung cancer [Lunca]	73
	NM_007294	BRCA1	-	DOID:219 / Colon cancer [Colca]	46
	NM_007294	BRCA1	-	DOID:10534 / Stomach cancer [Stoca]	38
	NM_007294	BRCA1	22197931	DOID:4159 / Skin cancer [Skica]	37

This chart displays the frequency of nsSNVs for the selected gene (y-axis) in the integrated BioMuta dataset for each cancer for which it is reported (x-axis). Taller bars indicate a greater frequency of variants in the selected gene for the corresponding cancer type. Note that the same nsSNV reported by multiple sources are counted as distinct variants for each source, and therefore frequencies reported in the chart do not reflect unique variants by position/substitution. Hovering the mouse over a bar will display a pop-up containing the cancer name and number of variants associated with that bar. Only cancers with non-zero counts for the queried gene are displayed in the chart.

Reported nsSNVs for P38398

This table lists all nsSNVs reported for P38398 across all cancer types.

Index	UniProtKB/SwissProt_ID	Gene_Name*	RefSeq_Nucleotide_ID	Genome_Position	Position_Nuc	Ref_Nuc	Var_Nuc	Position_AA	Ref_AA	Var_A
1	P38398	BRCA1	NM_007300	chr17:41276111-41276111	3	G	T	1	M	I
2	P38398	BRCA1	NM_007300	chr17:41276113-41276113	1	A	G	1	M	V
3	P38398	BRCA1	NM_007294	chr17:41276113-41276113	1	A	G	1	M	V
4	P38398	BRCA1	NM_007299	chr17:41276113-41276113	1	A	G	1	M	V
5	P38398	BRCA1	NM_007300	chr17:41276113-41276113	1	A	G	1	M	V
6	P38398	BRCA1	NM_007299	chr17:41276113-41276113	1	A	G	1	M	V
7	P38398	BRCA1	NM_007294	chr17:41276113-41276113	1	A	G	1	M	V
8	P38398	BRCA1	NM_007298	chr17:41276113-41276113	1	A	G	1	M	V
9	P38398	BRCA1	NM_007298	chr17:41276113-41276113	1	A	G	1	M	V
10	P38398	BRCA1	NM_007299	chr17:41276111-41276111	3	G	T	1	M	I
11	P38398	BRCA1	NM_007294	chr17:41276111-41276111	3	G	T	1	M	I
12	P38398	BRCA1	NM_007298	chr17:41276111-41276111	3	G	T	1	M	I
13	P38398	BRCA1	NM_007300	chr17:41276106-41276106	8	T	G	3	L	STOP
14	P38398	BRCA1	NM_007299	chr17:41276106-41276106	8	T	G	3	L	STOP
15	P38398	BRCA1	NM_007294	chr17:41276106-41276106	8	T	G	3	L	STOP
16	P38398	BRCA1	NM_007298	chr17:41276106-41276106	8	T	G	3	L	STOP
17	P38398	BRCA1	NM_007298	chr17:41276095-41276095	19	C	T	7	R	C
18	P38398	BRCA1	NM_007294	chr17:41276094-41276094	20	G	T	7	R	L
19	P38398	BRCA1	NM_007298	chr17:41276094-41276094	20	G	T	7	R	L
20	P38398	BRCA1	NM_007298	chr17:41276094-41276094	20	G	T	7	R	L
21	P38398	BRCA1	NM_007294	chr17:41276094-41276094	20	G	T	7	R	L
22	P38398	BRCA1	NM_007299	chr17:41276094-41276094	20	G	T	7	R	L
23	P38398	BRCA1	NM_007299	chr17:41276094-41276094	20	G	T	7	R	L
24	P38398	BRCA1	NM_007300	chr17:41276094-41276094	20	G	T	7	R	L

*Click to link to BioXpress database for related gene differential expression in human cancers

Statistics for BioMuta v3.0

- UniProtKB/Swiss-Prot Accessions hit: 18269
- Disease Ontology terms mapped: 76
- Number of nsSNVs reported in cancer: 4683912
- Number of predicted damaging mutations: 2304669
- Number of mutations affecting PTM sites: 980412

Downloads

BioMuta is updated every three months. Relevant subsets of data can be downloaded directly from the website by following the download link(s) on any search results page. To access complete dataset downloads for the current BioMuta version, please use the following links:

BioMuta Archive

Files of complete datasets corresponding to the current BioMuta version, as well as all available previous versions, can be downloaded from the links below. Please view the related READMEs below for more information.

APIs for BioMuta v3.0

The BioMuta API provides access to curated single-nucleotide variation (SNV) and disease association database where variations are mapped to the genomes, RefSeq nucleotide entries, and unified through UniProtKB/Swiss-Prot positional coordinates.

The base URL for the API is https://hive.biochemistry.gwu.edu/biomuta/api

Publication(s)

Wu TJ, Shamsaddini A, Pan Y, Smith K, Crichton DJ, Simonyan V and Mazumder R. A framework for organizing cancer-related variations from existing databases, publications and NGS data using a High-performance Integrated Virtual Environment (HIVE). Database. 2014. PMID: 24667251.

Overview

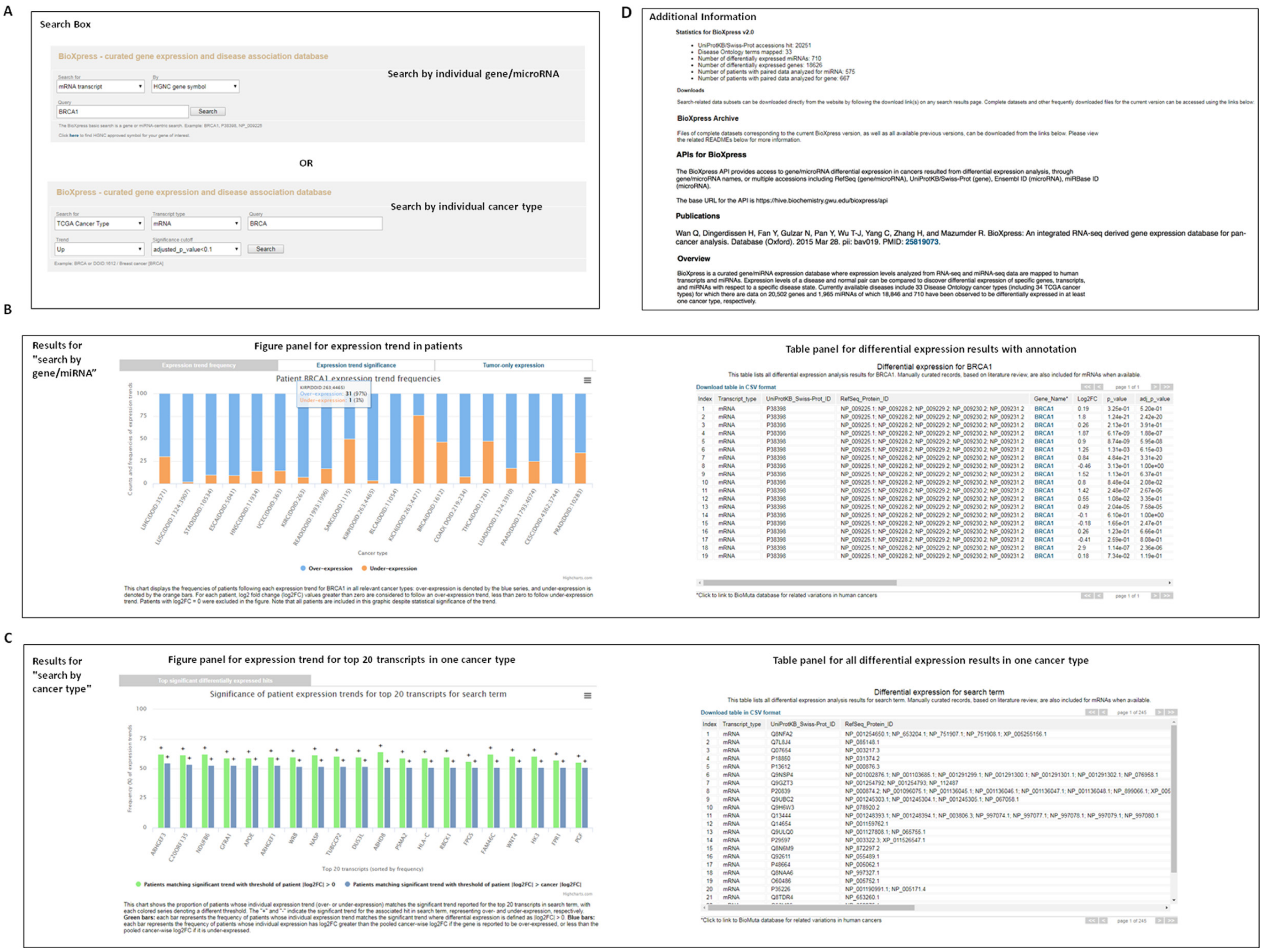
BioMuta is an annotated single-nucleotide variation (SNV) and disease association database where the variations are mapped to positions at the genome, protein, and gene levels. The current BioMuta database (BioMuta v3.0) contains a total of 6,845,387 nonsynonymous SNVs (nsSNVs) reported in 18,328 genes across 77 Disease Ontology cancer types. Additionally, a small portion of variants were added based on manual literature review of PubMed abstracts.

Figure 2. Interface of BioMuta. Two types of search engines are included in the interface of BioMuta (basic search for any gene name or accession, and advanced search for combined search of up to four search terms). After clicking 'search,' an interactive interim results page populates showing all possible genes matching the search criteria. After users select and click on the UniProtKB ID of their preferred gene, the detailed results page loads, displaying a figure and a table with information about all related SNVs and cancer types for this gene. APIs are integrated and users can obtain search results in JSON format by providing specific URLs. In addition to the search function, whole BioMuta datasets are downloadable, available from the tool home and archive pages.

BioMuta v3.0. The BioMuta basic search is a gene-centric search, requiring only the input of a single gene preferably in the format of HGNC approved gene symbol, UniProtKB accession, or RefSeq Gene ID. After submission of search term, an interim results page will load displaying all database hits to the specified query. From this page, the user can click the hyperlinked UniProtKB accession to navigate to the detailed results page for that particular accession record. Results pages have two primary components: charts and hit table. The default chart for BioMuta search displays the frequency of nsSNVs for the queried gene for each cancer type in which it is observed to be mutated. Hovertext displays additional information including the full cancer name, corresponding DOID, and variant count for the queried gene in that cancer. The second chart displays the frequency of nsSNVs for the queried gene along the length of the encoded protein. The hit table contains a variety of identifiers and annotations as described in the methods section above, and hyperlinks to relevant BioXpress entries and other external resources as appropriate.

The BioMuta advanced search allows the user to search specific terms by field and to combine search terms using AND/OR junctions to form a logical query. Up to four search terms may be combined to form such a logical query.

BioXpress v2.0. Similar to BioMuta, the BioXpress basic search is a gene or miRNA transcript-centric search where preferred search terms for mRNA search are HGNC approved symbol, UniProtKB accession or RefSeq gene IDs. For miRNA search, one can use miRNA alias, RefSeq, Ensembl, or miRBase accessions. Query submission will redirect the user directly to the results page, organized into the same basic chart and table components as described for BioMuta. The default chart for BioXpress transcript search displays the frequencies of patients following each expression trend (over- or under-expression) for the queried transcript across all relevant cancer types wherein patients with $\log_2$ fold change ($\log_2$FC) values greater than zero are considered to follow an over-expression trend, less than zero to follow an under-expression trend. The second chart shows the proportion of patients whose individual expression trend matches the significant trend reported for the queried transcript across different cancer types, with two colored series denoting different thresholds. The third chart shows a box plot of tumor sample expression, including expression from those samples with matched normal data and all unpaired tumor samples from TCGA. In addition to expression values and various statistics, the hit table contains a variety of identifiers and hyperlinks to relevant BioMuta entries and other external resources as appropriate.


A
Search Box
BioXpress - curated gene expression and disease association database
Search for
mRNA transcript ▾
By
HGNC gene symbol ▾
Query
BRCA1
Search
Search by individual gene/microRNA
OR
BioXpress - curated gene expression and disease association database
Search for
TCGA Cancer Type ▾
Transcript type
mRNA ▾
Query
BRCA
Trend
Up ▾
Significance cutoff
adjusted_p_value<0.1 ▾
Search
Search by individual cancer type
B
Results for "search by gene/miRNA"
Figure panel for expression trend in patients
Expression trend frequency
Expression trend significance
Tumor-only expression
Patient BRCA1 expression trend frequencies
KIRP(DOID:263:4465)
Over-expression: 31 (97%)
Under-expression: 1 (3%)
Counts and frequencies of expression trends
100
75
50
25
0
LIHC(DOID:3571)
LUSC(DOID:1324:3907)
STAD(DOID:10534)
ESCA(DOID:5041)
HNSC(DOID:11934)
UCEC(DOID:363)
KIRC(DOID:263)
READ(DOID:1993:1996)
SARC(DOID:1115)
KIRP(DOID:263:4465)
BLCA(DOID:11054)
KICH(DOID:263:4471)
BRCA(DOID:1612)
COAD(DOID:219:234)
THCA(DOID:1781)
LUAD(DOID:1324:3910)
PAAD(DOID:1793:4074)
CESC(DOID:4362:3744)
PRAD(DOID:10283)
Cancer type
● Over-expression
● Under-expression
Highcharts.com
Table panel for differential expression results with annotation
C
Results for "search by cancer type"
Figure panel for expression trend for top 20 transcripts in one cancer type
Top significant differentially expressed hits
Significance of patient expression trends for top 20 transcripts for search term
Frequency (%) of expression trends
100
75
50
25
0
ARHGEF3
C20ORF135
NDUFB6
GFRA1
APOE
ARHGEF1
WRB
NASP
TUBGCP2
DUS3L
ABHD8
PSMA2
HLA-C
RBCK1
FPGS
FAM46C
WNT4
HK3
FPR1
PGF
Top 20 transcripts (sorted by frequency)
● Patients matching significant trend with threshold of patient |log2FC| > 0
● Patients matching significant trend with threshold of patient |log2FC| > cancer |log2FC|
Highcharts.com
Table panel for all differential expression results in one cancer type
D

The BioXpress basic search is a gene or miRNA-centric search. Example: BRCA1, P38398, NP_009225

Click **here** to find HGNC approved symbol for your gene of interest.

Example: BRCA or DOID:1612 / Breast cancer [BRCA]

This chart displays the frequencies of patients following each expression trend for BRCA1 in all relevant cancer types: over-expression is denoted by the blue series, and under-expression is denoted by the orange bars. For each patient, log2 fold change (log2FC) values greater than zero are considered to follow an over-expression trend, less than zero to follow under-expression trend. Patients with log2FC = 0 were excluded in the figure. Note that all patients are included in this graphic despite statistical significance of the trend.

Differential expression for BRCA1

This table lists all differential expression analysis results for BRCA1. Manually curated records, based on literature review, are also included for mRNAs when available.

Download table in CSV format   << < page 1 of 1 > >>

Index	Transcript_type	UniProtKB_Swiss-Prot_ID	RefSeq_Protein_ID	Gene_Name*	Log2FC	p_value	adj_p_value
1	mRNA	P38398	NP_009225.1; NP_009228.2; NP_009229.2; NP_009230.2; NP_009231.2	BRCA1	0.19	3.25e-01	5.20e-01
2	mRNA	P38398	NP_009225.1; NP_009228.2; NP_009229.2; NP_009230.2; NP_009231.2	BRCA1	1.8	1.24e-21	2.42e-20
3	mRNA	P38398	NP_009225.1; NP_009228.2; NP_009229.2; NP_009230.2; NP_009231.2	BRCA1	0.26	2.13e-01	3.91e-01
4	mRNA	P38398	NP_009225.1; NP_009228.2; NP_009229.2; NP_009230.2; NP_009231.2	BRCA1	1.87	6.17e-09	1.88e-07
5	mRNA	P38398	NP_009225.1; NP_009228.2; NP_009229.2; NP_009230.2; NP_009231.2	BRCA1	0.9	8.74e-09	5.95e-08
6	mRNA	P38398	NP_009225.1; NP_009228.2; NP_009229.2; NP_009230.2; NP_009231.2	BRCA1	1.25	1.31e-03	6.15e-03
7	mRNA	P38398	NP_009225.1; NP_009228.2; NP_009229.2; NP_009230.2; NP_009231.2	BRCA1	0.84	4.84e-21	3.31e-20
8	mRNA	P38398	NP_009225.1; NP_009228.2; NP_009229.2; NP_009230.2; NP_009231.2	BRCA1	-0.46	3.13e-01	1.00e+00
9	mRNA	P38398	NP_009225.1; NP_009228.2; NP_009229.2; NP_009230.2; NP_009231.2	BRCA1	1.52	1.13e-01	6.37e-01
10	mRNA	P38398	NP_009225.1; NP_009228.2; NP_009229.2; NP_009230.2; NP_009231.2	BRCA1	0.8	8.48e-04	2.08e-02
11	mRNA	P38398	NP_009225.1; NP_009228.2; NP_009229.2; NP_009230.2; NP_009231.2	BRCA1	1.42	2.48e-07	2.67e-06
12	mRNA	P38398	NP_009225.1; NP_009228.2; NP_009229.2; NP_009230.2; NP_009231.2	BRCA1	0.55	1.08e-02	3.35e-01
13	mRNA	P38398	NP_009225.1; NP_009228.2; NP_009229.2; NP_009230.2; NP_009231.2	BRCA1	0.49	2.04e-05	7.58e-05
14	mRNA	P38398	NP_009225.1; NP_009228.2; NP_009229.2; NP_009230.2; NP_009231.2	BRCA1	-0.1	6.10e-01	1.00e+00
15	mRNA	P38398	NP_009225.1; NP_009228.2; NP_009229.2; NP_009230.2; NP_009231.2	BRCA1	-0.18	1.65e-01	2.47e-01
16	mRNA	P38398	NP_009225.1; NP_009228.2; NP_009229.2; NP_009230.2; NP_009231.2	BRCA1	0.26	1.23e-01	6.66e-01
17	mRNA	P38398	NP_009225.1; NP_009228.2; NP_009229.2; NP_009230.2; NP_009231.2	BRCA1	-0.41	2.59e-01	8.08e-01
18	mRNA	P38398	NP_009225.1; NP_009228.2; NP_009229.2; NP_009230.2; NP_009231.2	BRCA1	2.9	1.14e-07	2.36e-06
19	mRNA	P38398	NP_009225.1; NP_009228.2; NP_009229.2; NP_009230.2; NP_009231.2	BRCA1	0.18	7.34e-02	1.19e-01

*Click to link to BioMuta database for related variations in human cancers   << < page 1 of 1 > >>

This chart shows the proportion of patients whose individual expression trend (over- or under-expression) matches the significant trend reported for the top 20 transcripts in search term, with each colored series denoting a different threshold. The "+" and "-" indicate the significant trend for the associated hit in search term, representing over- and under-expression, respectively. **Green bars:** each bar represents the frequency of patients whose individual expression trend matches the significant trend where differential expression is defined as |log2FC| > 0. **Blue bars:** each bar represents the frequency of patients whose individual expression has log2FC greater than the pooled cancer-wise log2FC if the gene is reported to be over-expressed, or less than the pooled cancer-wise log2FC if it is under-expressed.

Differential expression for search term

This table lists all differential expression analysis results for search term. Manually curated records, based on literature review, are also included for mRNAs when available.

Download table in CSV format   << < page 1 of 245 > >>

Index	Transcript_type	UniProtKB_Swiss-Prot_ID	RefSeq_Protein_ID
1	mRNA	Q8NFA2	NP_001254650.1; NP_653204.1; NP_751907.1; NP_751908.1; XP_005255156.1
2	mRNA	Q7L8J4	NP_085148.1
3	mRNA	Q07654	NP_003217.3
4	mRNA	P18850	NP_031374.2
5	mRNA	P13612	NP_000876.3
6	mRNA	Q9NSP4	NP_001002876.1; NP_001103685.1; NP_001291299.1; NP_001291300.1; NP_001291301.1; NP_001291302.1; NP_076958.1
7	mRNA	Q9GZT3	NP_001254792; NP_001254793; NP_112487
8	mRNA	P20839	NP_000874.2; NP_001096075.1; NP_001136045.1; NP_001136046.1; NP_001136047.1; NP_001136048.1; NP_899066.1; XP_005
9	mRNA	Q9UBC2	NP_001245303.1; NP_001245304.1; NP_001245305.1; NP_067058.1
10	mRNA	Q9H6W3	NP_078920.2
11	mRNA	Q13444	NP_001248393.1; NP_001248394.1; NP_003806.3; NP_997074.1; NP_997077.1; NP_997078.1; NP_997079.1; NP_997080.1
12	mRNA	Q14654	NP_001159762.1
13	mRNA	Q9ULQ0	NP_001127808.1; NP_065755.1
14	mRNA	P29597	NP_003322.3; XP_011526547.1
15	mRNA	Q8N6M9	NP_872297.2
16	mRNA	Q92611	NP_055489.1
17	mRNA	P48664	NP_005062.1
18	mRNA	Q8NAA6	NP_997327.1
19	mRNA	O60486	NP_005752.1
20	mRNA	P35226	NP_001190991.1; NP_005171.4
21	mRNA	Q8TDR4	NP_653260.1

*Click to link to BioMuta database for related variations in human cancers   << < page 1 of 245 > >>

Additional Information

Statistics for BioXpress v2.0

- UniProtKB/Swiss-Prot accessions hit: 20251
- Disease Ontology terms mapped: 33
- Number of differentially expressed miRNAs: 710
- Number of differentially expressed genes: 18626
- Number of patients with paired data analyzed for miRNA: 575
- Number of patients with paired data analyzed for gene: 667

Downloads

Search-related data subsets can be downloaded directly from the website by following the download link(s) on any search results page. Complete datasets and other frequently downloaded files for the current version can be accessed using the links below:

BioXpress Archive

Files of complete datasets corresponding to the current BioXpress version, as well as all available previous versions, can be downloaded from the links below. Please view the related READMEs below for more information.

APIs for BioXpress

The BioXpress API provides access to gene/microRNA differential expression in cancers resulted from differential expression analysis, through gene/microRNA names, or multiple accessions including RefSeq (gene/microRNA), UniProtKB/Swiss-Prot (gene), Ensembl ID (microRNA), miRBase ID (microRNA).

The base URL for the API is https://hive.biochemistry.gwu.edu/bioxpress/api

Publications

Wan Q, Dingerdissen H, Fan Y, Gulzar N, Pan Y, Wu T-J, Yang C, Zhang H, and Mazumder R. BioXpress: An integrated RNA-seq derived gene expression database for pan-cancer analysis. Database (Oxford). 2015 Mar 28. pii: bav019. PMID: 25819073.

Overview

BioXpress is a curated gene/miRNA expression database where expression levels analyzed from RNA-seq and miRNA-seq data are mapped to human transcripts and miRNAs. Expression levels of a disease and normal pair can be compared to discover differential expression of specific genes, transcripts, and miRNAs with respect to a specific disease state. Currently available diseases include 33 Disease Ontology cancer types (including 34 TCGA cancer types) for which there are data on 20,502 genes and 1,965 miRNAs of which 18,846 and 710 have been observed to be differentially expressed in at least one cancer type, respectively.

Figure 3. Interface of BioXpress. Two types of search engines are included in the interface of BioXpress (search by individual gene/miRNA, and search by cancer type). For both searches, the results page for a queried gene/miRNA loads after clicking the 'search' button. In this page, a figure and a table displays information about either the expression trend of a single gene/miRNA in cancer patients (for gene/miRNA search) or the expression trend of the top 20 transcripts in one particular cancer type (for cancer type search), along with the results from differential expression analysis. APIs, whole dataset downloads, and documentation are also available.

Alternatively, the advanced cancer type search allows users to combine multiple search terms including cancer and transcript type in a single query, while simultaneously filtering hits by trend and significance threshold. The default view for the cancer type search shows the significance of patient expression for the top 20 transcripts matching the queried cancer type and meeting other search criteria.

DISCUSSION

Comparison to existing resources

Compared to several similar cancer or disease variant databases, BioMuta hosts a drastically greater number of variants due in part to the diversity of studies integrated into the single resource (Supplemental Table S4). The increased number of variants in BioMuta is also due to our policy of not strictly requiring literature citation, but rather using citation and other annotations as indicators of involvement in cancer. The number of samples in BioXpress is neither the greatest nor least among comparable resources, but it is important to note that all entries in BioXpress were generated by RNA-seq (not microarray) and represent tissue samples (not cell lines) (Supplemental Table S5). This is an intentional design element reflecting the paradigm shift toward RNA-seq strategies in the gene expression field. The greatest relative strength of both BioMuta and BioXpress is the unification by a central cancer disease ontology and the supplement of diverse functional annotations.

Please see Table 1 for more information.

Update features

New and updated data. Both previous versions of BioMuta and BioXpress were drafted prior to the completion of the TCGA project. We have now retrieved and analyzed a complete dump of all relevant data generated during the span of the TCGA project, resulting in more samples analyzed for certain cancer types and also containing certain entirely new cancer types. For BioXpress, the

Table 1. Summary and statistics of variants, expressed genes, and expressed miRNAs in current versions of BioMuta and BioXpress

BioMuta version/statistics	**BioMuta_v3.0**
Reviewed UniProtKB accessions hit	18 269
Disease Ontology terms mapped	77
Number of nsSNVs reported in cancer	4 684 236
Number of predicted damaging mutations	2 304 757
Number of mutations affecting PTM sites	980 447
BioXpress version/statistics	**BioXpress_v2.0**
Reviewed UniProtKB accessions hit	18 596
Disease Ontology terms mapped	33
Number of differentially expressed miRNAs	710
Number of differentially expressed genes	17 537
Number of patients with paired data analyzed for miRNA	575
Number of patients with paired data analyzed for gene	667

TCGA update alone has expanded the number of cancer types available for search from 27 to 33, and major effort was undertaken to include miRNA expression findings alongside the original updated set of mRNA expression. This involved substantial integration and mapping efforts to cross-map miRNA-seq data from TCGA and ICGC to HGNC, RefSeq, miRBase, Ensembl and Disease Ontology (DOID) terms as described above. For both resources, through a collaboration with a text-mining lab, we have also added over 1000 semi-automatically identified variants, with links to original publication PMIDs, and have a pilot set of similarly identified PMIDs for both mRNA and miRNA expression.

New pipeline for BioXpress. A substantial overhaul of the BioXpress pipeline including provision of QC measures has increased the quality and usability of this resource. We have re-analyzed data using a newer version of differential expression software (now DESeq2), and are planning to compare multiple analyses in future updates. We have defined a stricter set of criteria for interpretation of differential expression results, while making all analysis results (despite reported significance) available to the users for their own interpretation.

General usability. Although all entities are mapped to DOID terms, in this update, we also maintain the original terms provided by the primary data resources for the expression subset and allow search by either. All search options have expanded dramatically through the implementation of the BioMuta advanced search, allowing users to query essentially any field and value in the underlying database. Existing visualizations have been streamlined and updated, and in some cases new charts have been added to better represent database content and trends therein. Major updates to help documentation have occurred to enhance usability, and the backend has been optimized for ease of maintenance, scalability and performance.

Sustainability and future plans

Recently awarded funds will ensure the continued development of these resources for the next several years, as well as increasing exposure through collaborative research, primarily with members of the Early Detection Research Network (EDRN). This type of community-driven, sustainable development adheres to the mission of NCI's Informatics Technology for Cancer Research (ITCR), and is expected to result in provision of BioMuta and BioXpress as robust cancer variation and expression databases with active user engagement. For both resources, scripts are being streamlined and rewritten to allow easy maintenance and addition of new functionality, and interface pages will have embedded RDF. The next version of BioMuta is expected to expand to contain non-coding variants from whole genome sequencing experiments, and the next update to BioXpress will include additional RNA-seq samples as well as an integrated subset of BGEE (49) data. Pilots are actively underway to formalize a semi-automatic protocol to increase the number of literature citations in both databases.

CONCLUSIONS

As data continues to amass in volume, researchers will need to devise better ways to compute or digest, store, disseminate, and maintain it. Both BioMuta and BioXpress address several facets of this big data crunch by curating cancer related datasets covering SNVs and differential expression, making them consumable to a broad range of audiences, and engaging with users to guide development in a useful and sustainable approach. These datasets are compact, easy to use, have valuable annotations, and can be downloaded in bulk or accessed through web or API interfaces. Both knowledgebases can be used in complementary studies and represent valuable tool for cancer research.

DATA AVAILABILITY

BioMuta and BioXpress are hosted on the High-performance Integrated Virtual Environment (HIVE) server at the George Washington University, available through an open access web portal at http://hive.biochemistry.gwu.edu/biomuta and http://hive.biochemistry.gwu.edu/biomuta, respectively.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors would like to thank Dan Crichton for his assistance in linking to these resources from the EDRN portal, and Dr Vijay Shanker and Samir Gupta for their contribution and customization of literature mining tools. Authors would also like to thank Amanda Bell and Jeet Vora for testing and critical feedback.

FUNDING

National Institutes of Health [U01 CA215010]; EDRN Associate Member agreement [156620 to R.M.]. Funding for open access charge: McCormick Genomic and Proteomic Center, Department of Biochemistry and Molecular Medicine, George Washington University.
Conflict of interest statement. None declared.

REFERENCES

1. Welter,D., MacArthur,J., Morales,J., Burdett,T., Hall,P., Junkins,H., Klemm,A., Flicek,P., Manolio,T., Hindorff,L. *et al.* (2014) The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.*, **42**, D1001–D1006.
2. Collins,F.S., Guyer,M.S. and Charkravarti,A. (1997) Variations on a theme: cataloging human DNA sequence variation. *Science*, **278**, 1580–1581.
3. Ohashi,J. and Tokunaga,K. (2001) The power of genome-wide association studies of complex disease genes: statistical limitations of indirect approaches using SNP markers. *J. Hum. Genet.*, **46**, 478–482.
4. Makowsky,R., Pajewski,N.M., Klimentidis,Y.C., Vazquez,A.I., Duarte,C.W., Allison,D.B. and de los Campos,G. (2011) Beyond missing heritability: prediction of complex traits. *PLoS Genet.*, **7**, e1002051.
5. Speed,D. and Balding,D.J. (2014) MultiBLUP: improved SNP-based prediction for complex traits. *Genome Res.*, **24**, 1550–1557.
6. DeCotiis,J.L. and Lukac,D.M. (2017) KSHV and the role of notch receptor dysregulation in disease progression. *Pathogens*, **6**, E34.
7. Putteeraj,M., Yahaya,M.F. and Teoh,S.L. (2017) MicroRNA dysregulation in Alzheimer's disease. *CNS Neurol. Disord. Drug Targets*, doi:10.2174/1871527316666170807142311.
8. Fernandez-Calleja,V., Hernandez,P., Schvartzman,J.B., Garcia de Lacoba,M. and Krimer,D.B. (2017) Differential gene expression analysis by RNA-seq reveals the importance of actin cytoskeletal proteins in erythroleukemia cells. *PeerJ*, **5**, e3432.
9. Shen,Y., Bu,L., Li,R., Chen,Z., Tian,F., Lu,N., Ge,Q., Bai,Y. and Lu,Z. (2017) Screening effective differential expression genes for hepatic carcinoma with metastasis in the peripheral blood mononuclear cells by RNA-seq. *Oncotarget*, **8**, 27976–27989.
10. Sherry,S.T., Ward,M.H., Kholodov,M., Baker,J., Phan,L., Smigielski,E.M. and Sirotkin,K. (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.*, **29**, 308–311.
11. Stenson,P.D., Mort,M., Ball,E.V., Evans,K., Hayden,M., Heywood,S., Hussain,M., Phillips,A.D. and Cooper,D.N. (2017) The Human Gene Mutation Database: towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies. *Hum. Genet.*, **136**, 665–677.
12. Panwar,B., Omenn,G.S. and Guan,Y. (2017) miRmine: a database of human miRNA expression profiles. *Bioinformatics*, **33**, 1554–1560.
13. Clough,E. and Barrett,T. (2016) The Gene Expression Omnibus Database. *Methods Mol. Biol.*, **1418**, 93–110.
14. Landrum,M.J., Lee,J.M., Benson,M., Brown,G., Chao,C., Chitipiralla,S., Gu,B., Hart,J., Hoffman,D., Hoover,J. *et al.* (2016) ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Res.*, **44**, D862–D868.
15. Griffith,M., Spies,N.C., Krysiak,K., McMichael,J.F., Coffman,A.C., Danos,A.M., Ainscough,B.J., Ramirez,C.A., Rieke,D.T., Kujan,L. *et al.* (2017) CIViC is a community knowledgebase for expert crowdsourcing the clinical interpretation of variants in cancer. *Nat. Genet.*, **49**, 170–174.
16. Petryszak,R., Keays,M., Tang,Y.A., Fonseca,N.A., Barrera,E., Burdett,T., Fullgrabe,A., Fuentes,A.M., Jupp,S., Koskinen,S. *et al.* (2016) Expression Atlas update–an integrated database of gene and protein expression in humans, animals and plants. *Nucleic Acids Res.*, **44**, D746–D752.
17. Rhodes,D.R., Yu,J., Shanker,K., Deshpande,N., Varambally,R., Ghosh,D., Barrette,T., Pandey,A. and Chinnaiyan,A.M. (2004) ONCOMINE: a cancer microarray database and integrated data-mining platform. *Neoplasia*, **6**, 1–6.
18. Lee,P.H. and Shatkay,H. (2008) F-SNP: computationally predicted functional SNPs for disease association studies. *Nucleic Acids Res.*, **36**, D820–D824.
19. Ainscough,B.J., Griffith,M., Coffman,A.C., Wagner,A.H., Kunisaki,J., Choudhary,M.N., McMichael,J.F., Fulton,R.S., Wilson,R.K., Griffith,O.L. *et al.* (2016) DoCM: a database of curated mutations in cancer. *Nat. Methods*, **13**, 806–807.
20. Gao,J., Aksoy,B.A., Dogrusoz,U., Dresdner,G., Gross,B., Sumer,S.O., Sun,Y., Jacobsen,A., Sinha,R., Larsson,E. *et al.* (2013) Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci. Signal.*, **6**, pl1.
21. Cerami,E., Gao,J., Dogrusoz,U., Gross,B.E., Sumer,S.O., Aksoy,B.A., Jacobsen,A., Byrne,C.J., Heuer,M.L., Larsson,E. *et al.* (2012) The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov.*, **2**, 401–404.
22. Barretina,J., Caponigro,G., Stransky,N., Venkatesan,K., Margolin,A.A., Kim,S., Wilson,C.J., Lehar,J., Kryukov,G.V., Sonkin,D. *et al.* (2012) The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, **483**, 603–607.
23. Uhlen,M., Fagerberg,L., Hallstrom,B.M., Lindskog,C., Oksvold,P., Mardinoglu,A., Sivertsson,A., Kampf,C., Sjostedt,E., Asplund,A. *et al.* (2015) Proteomics. Tissue-based map of the human proteome. *Science*, **347**, 1260419.
24. Shin,G., Kang,T.W., Yang,S., Baek,S.J., Jeong,Y.S. and Kim,S.Y. (2011) GENT: gene expression database of normal and tumor tissues. *Cancer Inform.*, **10**, 149–157.
25. Zhang,J., Baran,J., Cros,A., Guberman,J.M., Haider,S., Hsu,J., Liang,Y., Rivkin,E., Wang,J., Whitty,B. *et al.* (2011) International Cancer Genome Consortium Data Portal—a one-stop shop for cancer genomics data. *Database (Oxford)*, **2011**, bar026.
26. Wu,T.J., Shamsaddini,A., Pan,Y., Smith,K., Crichton,D.J., Simonyan,V. and Mazumder,R. (2014) 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.
27. Pan,Y., Karagiannis,K., Zhang,H., Dingerdissen,H., Shamsaddini,A., Wan,Q., Simonyan,V. and Mazumder,R. (2014) Human germline and pan-cancer variomes and their distinct functional profiles. *Nucleic Acids Res.*, **42**, 11570–11588.
28. Wan,Q., Dingerdissen,H., Fan,Y., Gulzar,N., Pan,Y., Wu,T.J., Yan,C., Zhang,H. and Mazumder,R. (2015) BioXpress: an integrated RNA-seq-derived gene expression database for pan-cancer analysis. *Database (Oxford)*, **2015**, bav019.
29. Faison,W.J., Rostovtsev,A., Castro-Nallar,E., Crandall,K.A., Chumakov,K., Simonyan,V. and Mazumder,R. (2014) Whole genome single-nucleotide variation profile-based phylogenetic tree building methods for analysis of viral, bacterial and human genomes. *Genomics*, **104**, 1–7.
30. Wu,T.J., Schriml,L.M., Chen,Q.R., Colbert,M., Crichton,D.J., Finney,R., Hu,Y., Kibbe,W.A., Kincaid,H., Meerzaman,D. *et al.* (2015) Generating a focused view of disease ontology cancer terms for pan-cancer data integration and analysis. *Database (Oxford)*, **2015**, bav032.
31. Forbes,S.A., Beare,D., Gunasekaran,P., Leung,K., Bindal,N., Boutselakis,H., Ding,M., Bamford,S., Cole,C., Ward,S. *et al.* (2015) COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Res.*, **43**, D805–D811.
32. Gonzalez-Perez,A., Perez-Llamas,C., Deu-Pons,J., Tamborero,D., Schroeder,M.P., Jene-Sanz,A., Santos,A. and Lopez-Bigas,N. (2013) IntOGen-mutations identifies cancer drivers across tumor types. *Nat. Methods*, **10**, 1081–1082.
33. Pundir,S., Martin,M.J. and O'Donovan,C. (2017) UniProt Protein Knowledgebase. *Methods Mol. Biol.*, **1558**, 41–55.
34. Wang,K., Li,M. and Hakonarson,H. (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.*, **38**, e164.
35. O'Leary,N.A., Wright,M.W., Brister,J.R., Ciufo,S., Haddad,D., McVeigh,R., Rajput,B., Robbertse,B., Smith-White,B., Ako-Adjei,D.

et al. (2016) Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.*, **44**, D733–D745.

36. Tatusova,T., DiCuccio,M., Badretdin,A., Chetvernin,V., Nawrocki,E.P., Zaslavsky,L., Lomsadze,A., Pruitt,K.D., Borodovsky,M. and Ostell,J. (2016) NCBI prokaryotic genome annotation pipeline. *Nucleic Acids Res.*, **44**, 6614–6624.
37. Huang,H., McGarvey,P.B., Suzek,B.E., Mazumder,R., Zhang,J., Chen,Y. and Wu,C.H. (2011) A comprehensive protein-centric ID mapping service for molecular data integration. *Bioinformatics*, **27**, 1190–1191.
38. Sievers,F., Wilm,A., Dineen,D., Gibson,T.J., Karplus,K., Li,W., Lopez,R., McWilliam,H., Remmert,M., Soding,J. *et al.* (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.*, **7**, 539.
39. Adzhubei,I.A., Schmidt,S., Peshkin,L., Ramensky,V.E., Gerasimova,A., Bork,P., Kondrashov,A.S. and Sunyaev,S.R. (2010) A method and server for predicting damaging missense mutations. *Nat. Methods*, **7**, 248–249.
40. Marchler-Bauer,A., Bo,Y., Han,L., He,J., Lanczycki,C.J., Lu,S., Chitsaz,F., Derbyshire,M.K., Geer,R.C., Gonzales,N.R. *et al.* (2017) CDD/SPARCLE: functional classification of proteins via subfamily domain architectures. *Nucleic Acids Res.*, **45**, D200–D203.
41. Li,H., Xing,X., Ding,G., Li,Q., Wang,C., Xie,L., Zeng,R. and Li,Y. (2009) SysPTM: a systematic resource for proteomic research on post-translational modifications. *Mol. Cell. Proteomics*, **8**, 1839–1849.
42. Hornbeck,P.V., Chabra,I., Kornhauser,J.M., Skrzypek,E. and Zhang,B. (2004) PhosphoSite: a bioinformatics resource dedicated to physiological protein phosphorylation. *Proteomics*, **4**, 1551–1561.
43. Dinkel,H., Chica,C., Via,A., Gould,C.M., Jensen,L.J., Gibson,T.J. and Diella,F. (2011) Phospho.ELM: a database of phosphorylation sites–update 2011. *Nucleic Acids Res.*, **39**, D261–D267.
44. Lee,T.Y., Chen,Y.J., Lu,C.T., Ching,W.C., Teng,Y.C. and Huang,H.D. (2012) dbSNO: a database of cysteine S-nitrosylation. *Bioinformatics*, **28**, 2293–2295.
45. Keshava Prasad,T.S., Goel,R., Kandasamy,K., Keerthikumar,S., Kumar,S., Mathivanan,S., Telikicherla,D., Raju,R., Shafreen,B., Venugopal,A. *et al.* (2009) Human Protein Reference Database–2009 update. *Nucleic Acids Res.*, **37**, D767–D772.
46. Kibbe,W.A., Arze,C., Felix,V., Mitraka,E., Bolton,E., Fu,G., Mungall,C.J., Binder,J.X., Malone,J., Vasant,D. *et al.* (2015) Disease Ontology 2015 update: an expanded and updated database of human diseases for linking biomedical knowledge through disease data. *Nucleic Acids Res.*, **43**, D1071–D1078.
47. Zhu,Y., Qiu,P. and Ji,Y. (2014) TCGA-assembler: open-source software for retrieving and processing TCGA data. *Nat. Methods*, **11**, 599–600.
48. Love,M.I., Huber,W. and Anders,S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*, **15**, 550.
49. Bastian,F., Parmentier,G., Roux,J., Moretti,S., Laudet,V. and Robinson-Rechavi,M. (2008) Bgee: integrating and comparing heterogeneous transcriptome data among species. *Data Integration in the Life Sciences*, Springer, Berlin; Heidelberg, Vol. **5109**, pp. 124–131.