Database, 2021, 1–10
doi:10.1093/database/baab018
Database tool

Database tool

BC-TFdb: a database of transcription factor drivers in breast cancer

Abbas Khan[1], Taimoor Khan[1], Syed Nouman Nasir[2], Syed Shujait Ali[2], Muhammad Suleman[2], Muhammad Rizwan[2], Muhammad Waseem[3], Shahid Ali[2], Xia Zhao[4,*] and Dong-Qing Wei[1,5,6,*]

[1]Department of Bioinformatics and Biological Statistics, School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai 200240, P.R. China, [2]Center for Biotechnology and Microbiology, University of Swat, Swat, KP 19200, Pakistan, [3]Faculty of Rehabilitation and Allied Health Science, Riphah International University, Islamabad 44000, Pakistan, [4]Department of Microbiology, Army Medical University, Chongqing 400044, P.R. China, [5]State Key Laboratory of Microbial Metabolism, Shanghai-Islamabad-Belgrade Joint Innovation Center on Antibacterial Resistances, Joint International Research Laboratory of Metabolic and Developmental Sciences and School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai 200240, P.R. China and [6]Peng Cheng Laboratory, Vanke Cloud City Phase I Building 8, Xili Street, Nanshan District, Shenzhen, Guangdong 518055, P.R. China

*Corresponding author: Tel: 0086-13918500529; Fax: 862134204573; Email: dqwei@sjtu.edu.cn

Correspondence may also be addressed to Xia Zhao, Tel: 0086-13667652559; Fax: 02087343088.

Citation details: Khan, A., Khan, T., Nasir, S.N. *et al.* BC-TFdb: a database of transcription factor drivers in breast cancer. *Database* (2021) Vol. 2021: article ID baab018; doi:10.1093/database/baab018

Received 29 January 2021; Revised 1 March 2021; Accepted 9 April 2021

Abstract

Transcription factors (TFs) are DNA-binding proteins, which regulate many essential biological functions. In several cancer types, TF function is altered by various direct mechanisms, including gene amplification or deletion, point mutations, chromosomal translocations, expression alterations, as well as indirectly by non-coding DNA mutations influencing the binding of the TF. TFs are also actively involved in breast cancer (BC) initiation and progression. Herein, we have developed an open-access database, BC-TFdb (Breast Cancer Transcription Factors database), of curated, non-redundant TF involved in BC. The database provides BC driver TFs related information including genomic sequences, proteomic sequences, structural data, pathway information, mutations information, DNA binding residues, survival and therapeutic resources. The database will be a useful platform for researchers to obtain BC-related TF–specific information. High-quality datasets are downloadable for users to evaluate and develop computational methods for drug designing against BC.

Database URL: https://www.dqweilab-sjtu.com/index.php.

Introduction

Transcriptional regulation of genes is a vital process controlled by transcription factors (TFs) as key regulators to maintain cell homeostasis (1, 2). TFs regulate targeted gene expression by recognition and binding to specific DNA sequences known as transcription factor–binding sites (TFBSs) (3, 4). On the other hand, disruption of TF–target regulation leads to cellular damage and ultimately causes disease (4). Similarly, genetic alterations in cancer produce distinct tumor populations, which may remain benign or metastasize to distal sites (5, 6). For instance, heterogenic breast cancer (BC) tumor subtypes are underlined by a unique set of oncogenic alterations characterized in abnormal proliferation, invasion and metastatic potential (7). These genetic alterations are also co-related with drug resistance and relapse and offer attractive targets in therapeutic implications (8, 9).

TFs are also characterized in BC as tumor suppressors or oncogenes (10) and are key players in causing abnormal cellular growth (11). TFs also serve as important prognostic markers in BC. TF KLF4 (Kruppel-like factor 4) was reported as predictive pathological BC remission marker following neoadjuvant chemotherapy (12). Moreover, TFEts-1 expression was considered as an independent prognostic marker for recurrence-free survival (RFS) in BC (13). There is an increased interest in the identification of TFs as effective predictors for BC prognosis. Furthermore, it is also evident that tumor biomarker signature is crucial to explore more effective treatments for BC (8). Meanwhile, systematic studies to comprehensively visualize TF–target can be helpful to depict TF–target regulations. Integrated omics datasets for TFs (including TFBSs, target prediction, mRNA profiling and epigenetic status of chromatin) are considered as a useful resource for understanding TF–target regulations (14) and benefit researchers in transcriptional regulation studies (8).

In this study, we developed a database named BC-TFdb by utilizing a comprehensive approach to gather BC-related TF–target relationships. The resulting datasets in our study provide a comprehensive platform for studies related to TF–target regulation in BC. The BC-TFdb has integrated specific datasets with information from basic to advanced, reliable information to study TF–target regulations in BC. It will be helpful in providing online details to forecast potential co-association and co-regulation between TFs. In summary, BC-TFdb will serve as a useful resource for researchers to analyze TF regulation and gene expression in BC.

Methodology

Data collection and analysis

Bioinformatics tools have been widely utilized in large datasets to landscape genetic alterations in BC. Many groups identified these alterations through exome sequencing, copy number variation, mRNA and proteomic analyses of thousands of BC samples covering all major subtypes (9, 15, 16). Recently, whole-genome sequencing data of identified cancer-driving genetic alterations in 93 genes (17) and dataset of TFs relevant to RFS were also profiled (18). For database development, we collected human TFs associated with BC and involved in different BC pathways from different literatures (19, 20). Basic information regarding these TFs was also extracted from literature and used to divide into TF types. Genomic sequences (gene nucleotide sequences) and protein sequences of the collected set of TFs were retrieved from NCBI (National Center for Biotechnology Information) (https://www.ncbi.nlm.nih.gov/) (21) and UniProt platform (https://www.uniprot.org/) (22). Similarly, 3D structures for all data were downloaded from the PDB (Protein Data Bank) database (https://www.rcsb.org/) (23), and all the unavailable structures were submitted to Phyre2.0 for computational modeling (22, 24, 25). Information on the association of TFs in different BC-related pathways was obtained from KEGG (Kyoto Encyclopedia of Genes and Genomes) tool (26). In genetics, a missense mutation is a point mutation in which a single nucleotide change results in a codon that codes for a different amino acid. To obtain information on the missense mutations, the dbSNP (Single Nucleotide Polymorphism Database) database was accessed, and a list of mutations for each TF was retrieved. For screening these mutations, different tools such as SIFT (Scale-invariant feature transform) (27), PolyPhen (Polymorphism Phenotyping) (28) and CADD (Combined Annotation Dependent Depletion) (29) were utilized for prediction of deleterious nsSNPs (non-synonymous single nucleotide polymorphisms). SIFT tool (https://sift.bii.a-star.edu.sg/) determines the effect of amino acid substitution on protein's function based on its physicochemical properties. SIFT scores nsSNPs as deleterious when the score is between 0 and 0.05, while nsSNPs with a score between 0.05 and 1 are declared as tolerated by SIFT (30). PolyPhen (http://genetics.bwh.harvard.edu/pph2/) is a web tool that forecasts the likely impact of an amino acid replacement on the structure and function of a human protein by combining physical and comparative parameters. Further, we also used CADD (https://cadd.gs.washington.edu/) to predict the most deleterious SNPs for each protein (TF) included. This framework integrates multiple annotations into one metric by contrasting variants that survived natural selection with simulated mutations. To understand the role of each TF in the initiation, progression and post-prognostic value survival analysis was performed using Kaplan–Meier plotter (31). Kaplan–Meier estimate is one of the best options to be used to measure the fraction of subjects living for a certain amount of time after treatment. Next, we also

searched for the DNA-binding residues and designed a unique dataset included in our study design (32). For this purpose, we used DP-Bind (http://lcg.rit.albany.edu/dp-bind), which is a sequence-based approach for predicting the nucleic acid–binding residues in DNA-binding proteins (32). The server uses sequence evolutionary conservation information to predict the DNA-binding residues and their probability score. Finally, to address therapeutic implications, all the available drugs from PubChem (33) and DrugBank (34) were listed and included as a separate dataset for each TFs. Together these data will make a comprehensive platform for BC treatment options.

Development of database

Database development is an intricate task, as the designer has to accommodate the information that can be processed quickly, stored and responsive to the user needs easily. BC-TFdb offers large cloud-based online portals to ease data access and for in-depth analysis. BC-TFdb provides a friendly interface, allowing easy access to and efficient use of information. BC-TFdb search index was designed through various programming tools such as MySQL (Structured Query Language) (35), PHP (Hypertext Preprocessor) (36), AJAX (Asynchronous JavaScript And XML) (37), HTML (Hypertext Markup Language) (38), jQuery (39) and Bootstrap.

Back end preparation

Database designing was done using WAMP server. Open access was ensured by scripting in environments like HTML and PHP. Data storage, manipulation and retrieving from the databases were managed through MySQL. WAMP server provides an environment to create a web application with PHP, Apache and MySQL database and is equipped with PHP MyAdmin to confer full control over the web contents.

Front-end preparation

The database interface needs intelligence as the users interact directly with it and increase its overall acceptance. Client-side or front-end design aims to deliver an attractive interface to accomplish tasks efficiently and easily without having much of bioinformatics knowledge of command-line interface. Popular tools such as Bootstrap, CSS (Cascading Style Sheets) and HTML were used to tailor BC-TFdb. CSS is a high-level programming language commonly used to customize and style PHP and HTML script. The database provides a user-friendly platform for easy access and operation of TFs related data. The overall workflow of the work, including the data source and other information, is given in Figure 1.

Results and discussion

BC-TFdb primarily focuses on the multiple features regarding the TFs in BC. This database is a collective platform for a total of 161 TFs involved in BC progression. The database includes multiple tabs including basic information, sequence and structural information, pathway information, survival information, DNA-binding residues, missense mutations and their impact, and therapeutics for the representation of a specific type of data. The overall workflow of the strategy used in the design of this database has been given in Figure 1.

Structure and implementation of the database modules

Different modules have been used for designing the BC-TF-specific database and are discussed as follows.

Basic information tab

The basic information page represents a sheet that contains details about the genes and proteins, UniProt entry ID, organism type, length, KEGG accession IDs and string PPI (protein-protein interactions) accession number. This tab is also equipped with the search module for accessing any specific information about the TF of interest. This information will help the user to access specific information about each TF. The representative image of this tab is presented in Figure 2.

Sequence and structural information tab

Genomics, proteomics and structural information retrieved from UniProt, NCBI, RCSB (Research Collaboratory for Structural Bioinformatics), Ensembl and gnomAD are all available in this tab. This tab is further divided into three sub-modules, which represent gene sequences, protein sequences and structural information separately. The first sub-tab represents the gene sequences along with the gene names and nucleotide lengths. Similarly, the proteomics tab represents the protein sequences with the amino acid length and protein name. Users can access specific information on both genomics and proteomics by searching using the sequence features or gene name in the search engine. Additionally, the structural information tab provides data on these TFs. This sheet provides information about the gene name, PDB entry ID (RCSB), structural determination method (X-ray, NMR (Nuclear magnetic resonance), cryo-EM (Electron Microscopy) or homology model), resolution of the structure and chain information. This sheet also owns information about the AA position of the specific structure. Moreover, the download module

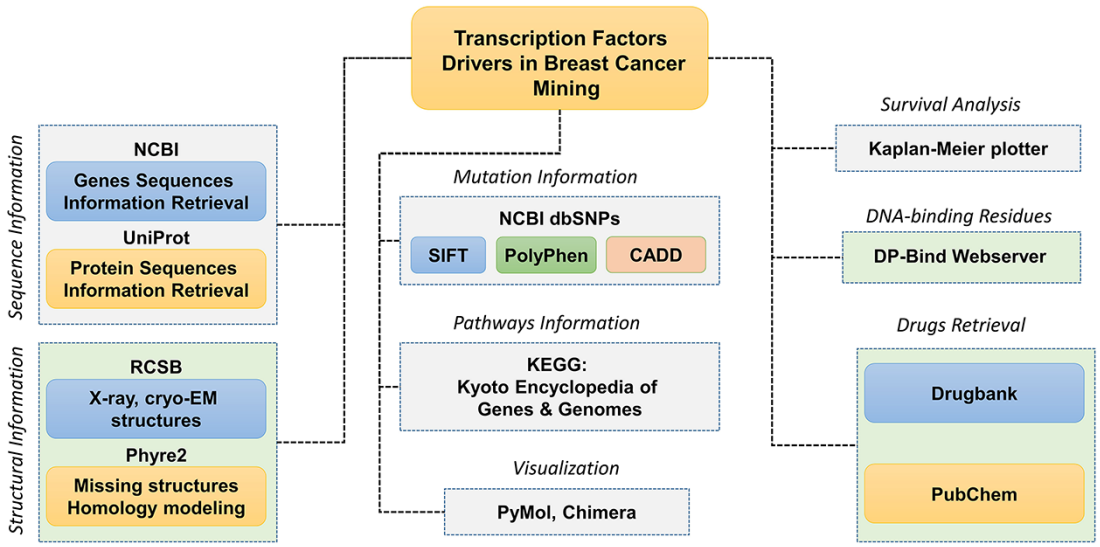


Figure 1. Schematic representation of the database components and data analysis modules used to obtain the specific results. Each tab is tagged, and their respective analyses of webservers or software are also given.

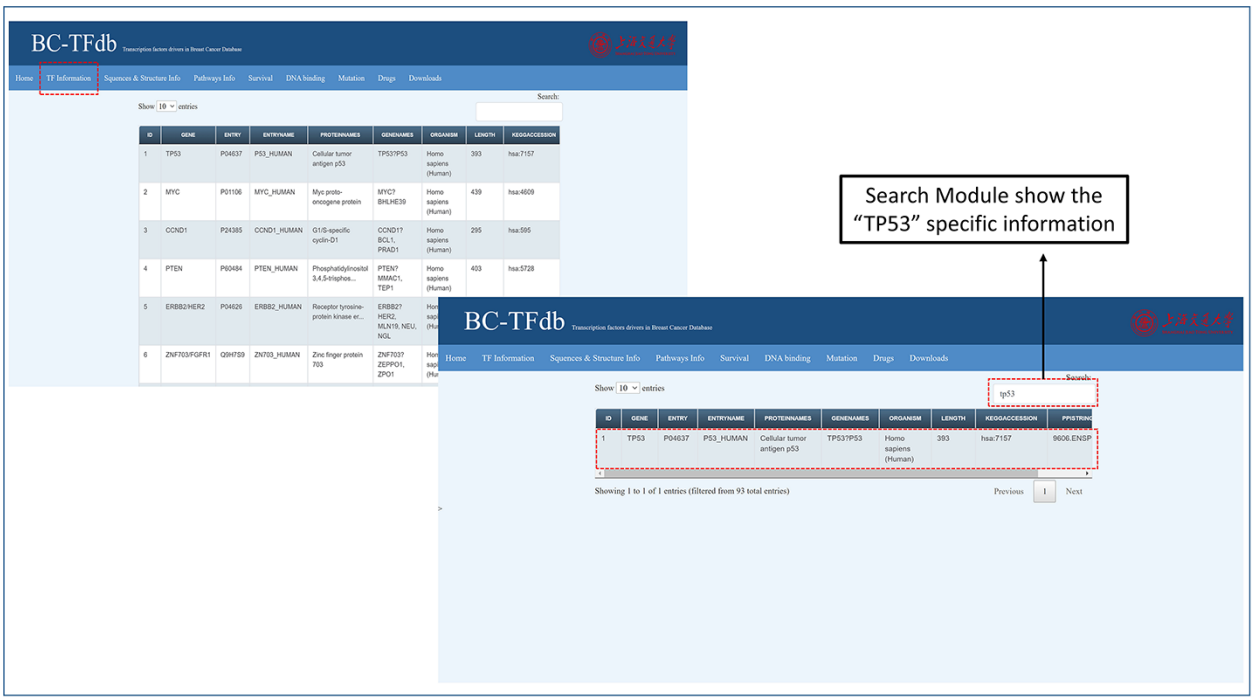


Figure 2. Representative image of the basic information tab showing details about TFs from the dataset of the cohort study. The bottom panel shows the search module implemented in the database to obtain specific information.

allows users to easily download the specific structure. Furthermore, 14 unavailable structures were modeled using Phyre2 homology modeling server and included as downloadable information deposited in the online database. The representative image tutorial of this tab is presented in Figure 3.

Pathway information tab

For pathway information about specific TFs, KEGG was accessed to retrieve listed information about the involvement of the TFs in BC pathways. This information was deposited on the pathway information tab for easy access to users. Specific pathway information can be

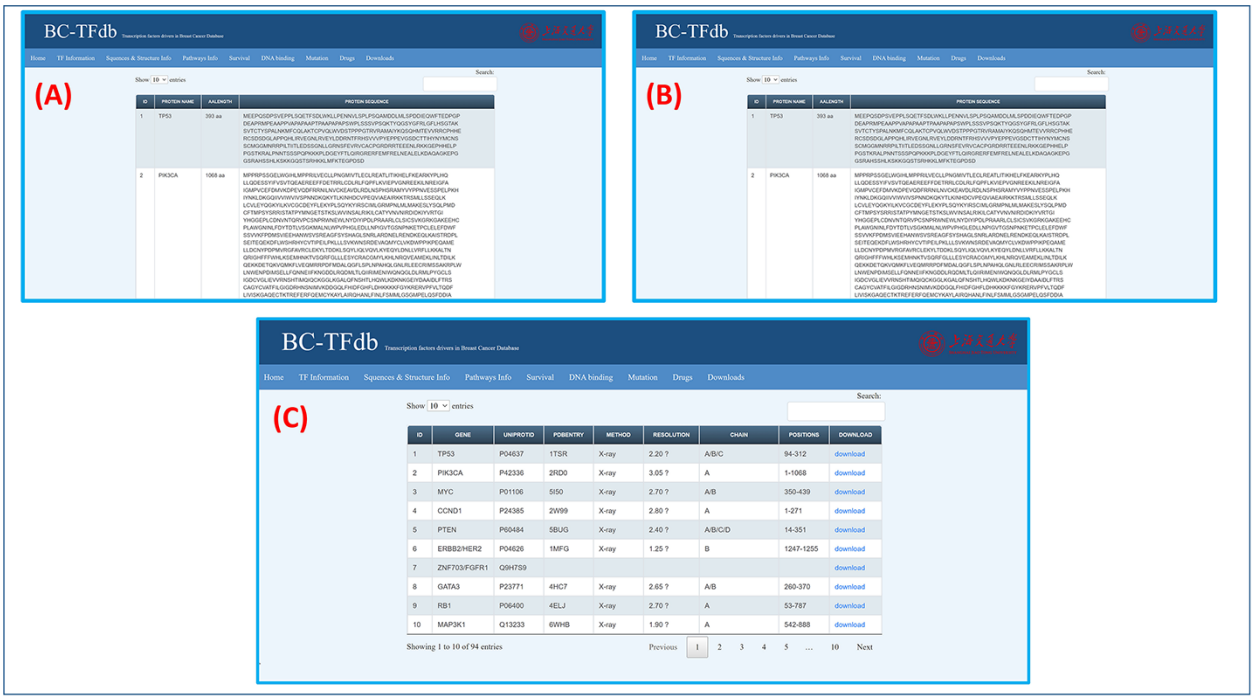


Figure 3. Representation of the Sequence and structural information tab. (A) shows the tutorial image for accessing the gene sequence information of TFs. (B) shows the tutorial image of accessing the protein sequence information of TFs. (C) shows the tutorial image of accessing the protein structural information of TFs.

viewed and downloaded as an image file. Further, the search module can also be utilized to retrieve the pathway-specific information based on the gene name. The image tutorial of the pathway information tab is represented in Figure 4.

Survival information tab

The database also includes survival information about each TF included in the study related to BC. Users can easily retrieve the TF-specific survival information such as *P* value and hazard ratio. Moreover, the sheet also has the gene name and Affymetrix ID of each TF. The specific information can be searched in the search module either using the gene name or Affymetrix ID. Furthermore, with the downloads module, user can also download the survival plot of each TF available in this tab. KM plotter was utilized for the survival analysis of BC-specific TFs, and survival plots for each of the TF were searched, retrieved and deposited here. The image tutorial of the Survival information tab is represented in Figure 5.

DNA binding tab

Information about the DNA-binding residues is crucial to understand the role of specific residues in the recognition of DNA or RNA by the TFs, which may help in the BC diagnosis and therapeutic development. These hotspot residues may also serve as drug hotspots for binding interactions and novel drug development. On the other hand, lack of such information hinders the exploration of the molecular mechanism of DNA or RNA recognition. It, thus, limits the understanding of cancer prognosis and development at a molecular level. No single platform is available to provide information about the role of specific residue in the DNA or RNA recognition. Thus, we used DP-Bind to predict the DNA-binding residues of each TF. The server predicts DNA-binding residue represented with 1 for binding and 0 for nonbinding residues. Herein, information about the DNA-binding residues of each TF is given. Only residues that were predicted to be DNA binding are given in the sheet. A total of 6180 residues for all the TFs are predicted to be DNA binding, while the rest were discarded as nonbinding residues. The DNA binding tab provides information about the protein name, residue position, residue name, SVM (support vector machine) binding score and SVM probability. The DNA-binding residues of each TF are arranged according to the highest SVM probability values. The search module using the gene name displays gene-specific DNA-binding residue information. Further, these analyses would also help the experimental researchers to validate the binding probability of each residue for a specific TF. The image tutorial of the DNA binding tab is represented in Figure 6.

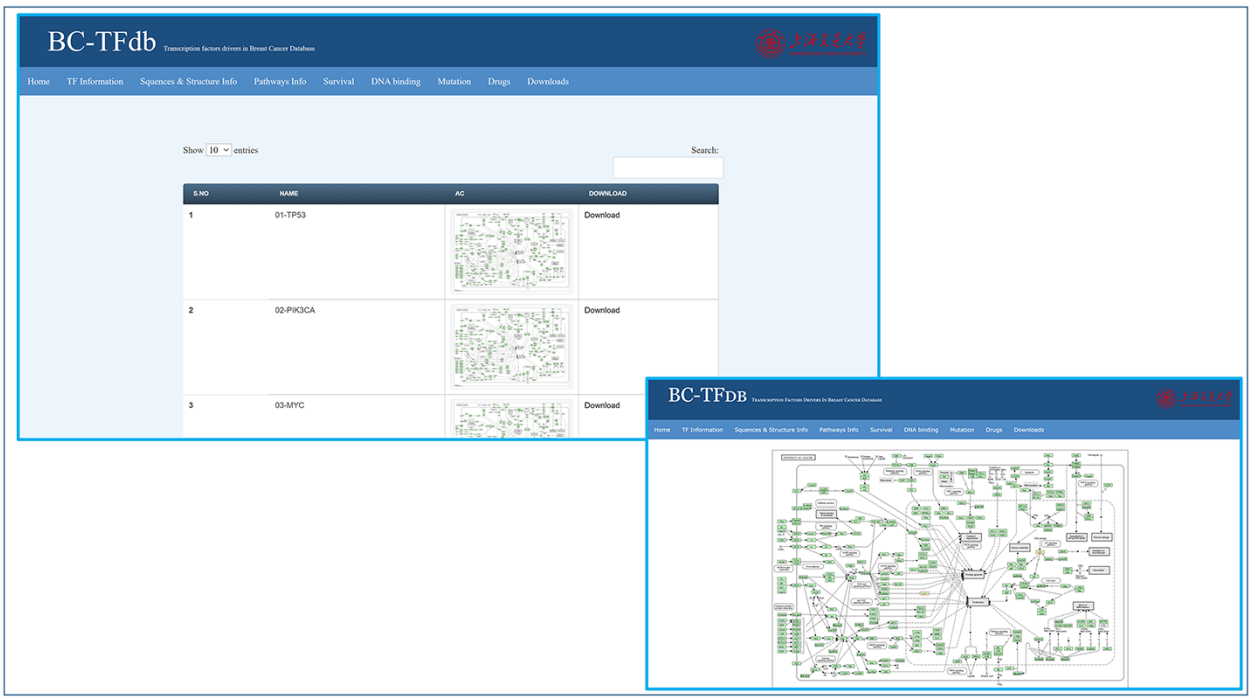


Figure 4. Representative tutorial image of the Pathway information tab upon online access. The bottom panel views a sample pathway of TP53.

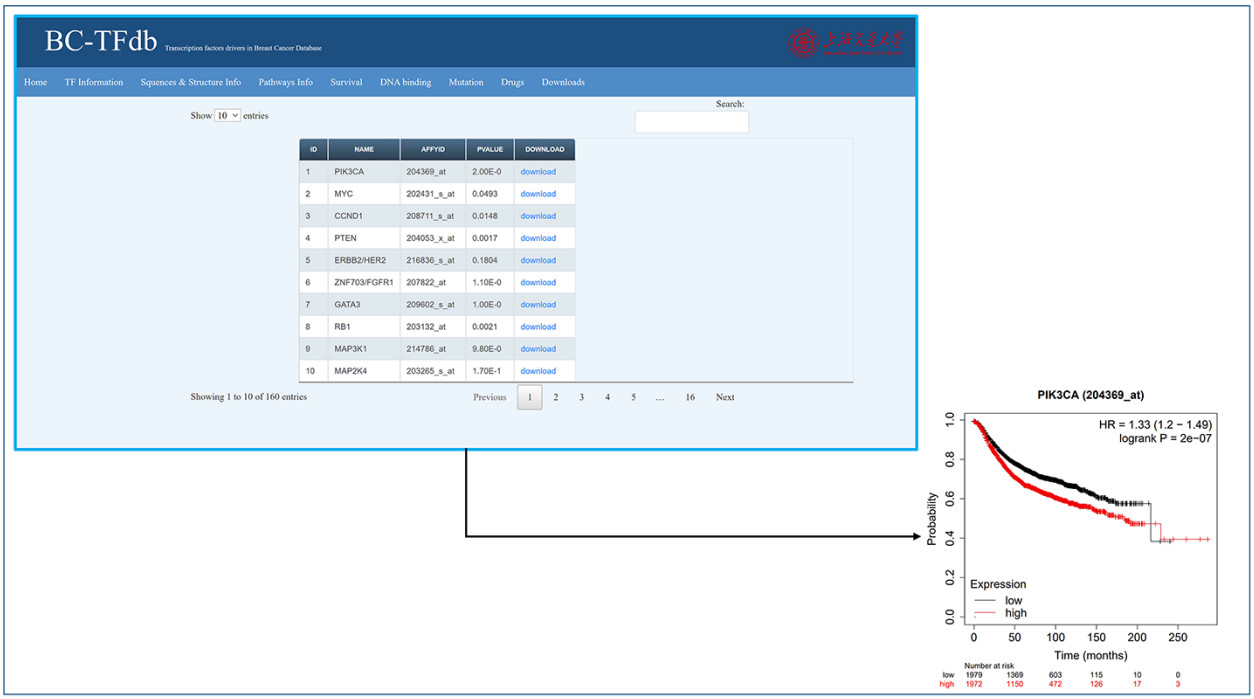
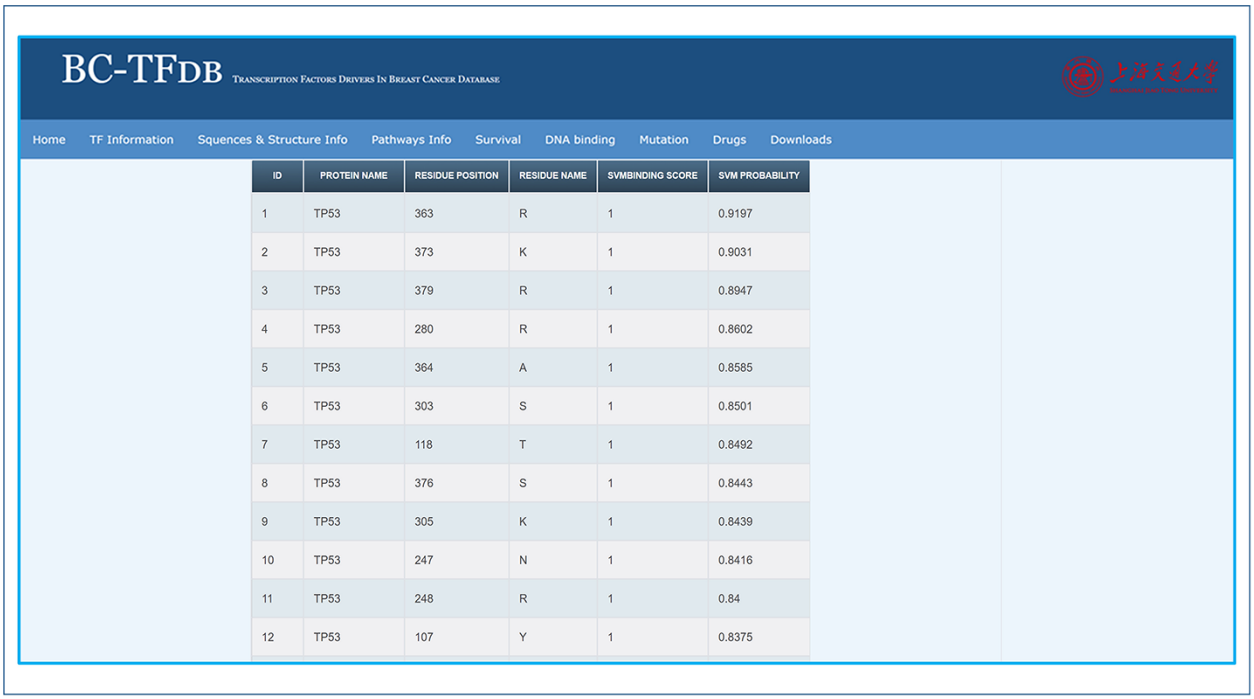


Figure 5. Representative image of the Survival information tab, including a list of all TFs and their downloadable KM (kaplan-meier plot) plots. The bottom panel shows the sample survival KM plot for PIK3CA.

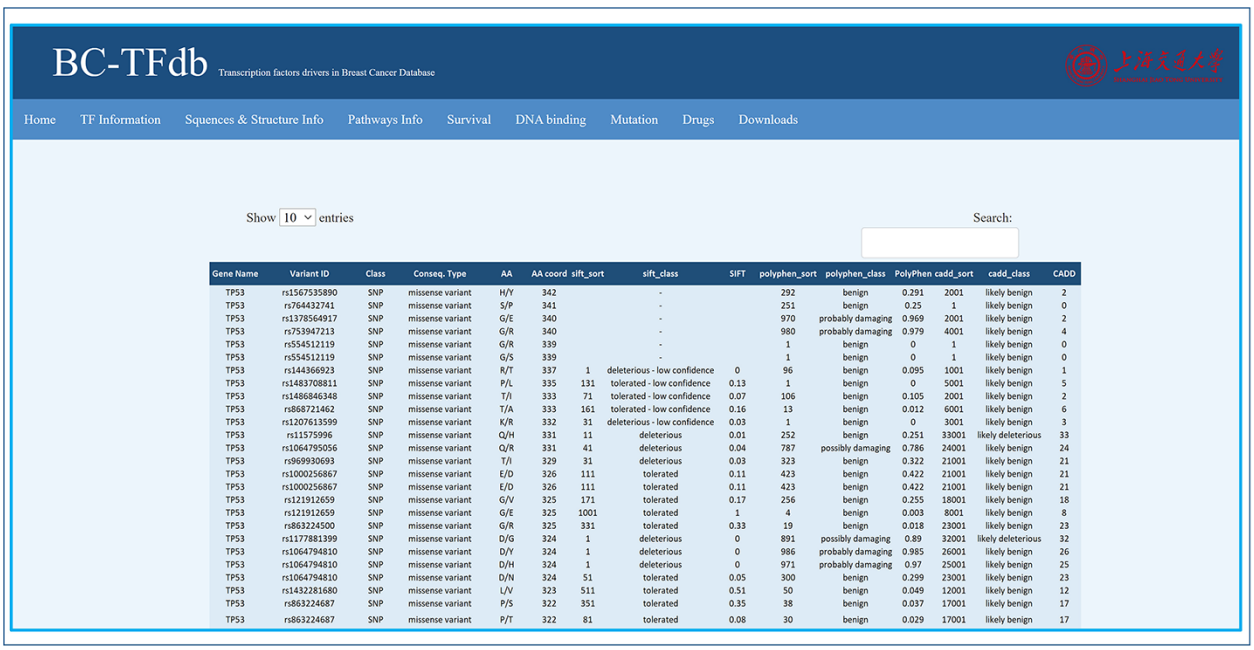
Mutations tab

Mutations are highly correlated with the complex disease phenotype and, thus, of great concern in cancer research. Any substitution in the protein-coding region may alter the protein structure and function. Therefore, depicting the impact of each mutation may help to understand the disease initiation and progression. Similarly, TF-specific mutations contribute as the driving force in BC progression. Moreover, the identification and characterization of these mutations are crucial to better understand the underlying



ID	PROTEIN NAME	RESIDUE POSITION	RESIDUE NAME	SVMBINDING SCORE	SVM PROBABILITY
1	TP53	363	R	1	0.9197
2	TP53	373	K	1	0.9031
3	TP53	379	R	1	0.8947
4	TP53	280	R	1	0.8602
5	TP53	364	A	1	0.8585
6	TP53	303	S	1	0.8501
7	TP53	118	T	1	0.8492
8	TP53	376	S	1	0.8443
9	TP53	305	K	1	0.8439
10	TP53	247	N	1	0.8416
11	TP53	248	R	1	0.84
12	TP53	107	Y	1	0.8375

Figure 6. Representative image of the DNA binding tab, including a list of all TFs and listed SVM scores.



Gene Name	Variant ID	Class	Conseq. Type	AA	AA coord	sift_sort	sift_class	SIFT	polyphen_sort	polyphen_class	PolyPhen	cadd_sort	cadd_class	CADD
TP53	rs1567535890	SNP	missense variant	H/Y	342		-		292	benign	0.291	2001	likely benign	2
TP53	rs764432741	SNP	missense variant	S/P	341		-		251	benign	0.25	1	likely benign	0
TP53	rs1378564917	SNP	missense variant	G/E	340		-		970	probably damaging	0.969	2001	likely benign	2
TP53	rs753947213	SNP	missense variant	G/R	340		-		980	probably damaging	0.979	4001	likely benign	4
TP53	rs554512119	SNP	missense variant	G/R	339		-		1	benign	0	1	likely benign	0
TP53	rs554512119	SNP	missense variant	G/S	339		-		1	benign	0	1	likely benign	0
TP53	rs144366923	SNP	missense variant	R/T	337	1	deleterious - low confidence	0	96	benign	0.095	1001	likely benign	1
TP53	rs1483708811	SNP	missense variant	P/L	335	131	tolerated - low confidence	0.13	1	benign	0	5001	likely benign	5
TP53	rs1486846348	SNP	missense variant	T/I	333	71	tolerated - low confidence	0.07	106	benign	0.105	2001	likely benign	2
TP53	rs868721462	SNP	missense variant	T/A	333	161	tolerated - low confidence	0.16	13	benign	0.012	6001	likely benign	6
TP53	rs1207613599	SNP	missense variant	R/R	332	31	deleterious - low confidence	0.03	1	benign	0	3001	likely benign	3
TP53	rs11575996	SNP	missense variant	Q/H	331	11	deleterious	0.01	252	benign	0.251	33001	likely deleterious	33
TP53	rs1064795056	SNP	missense variant	Q/R	331	41	deleterious	0.04	787	possibly damaging	0.786	24001	likely benign	24
TP53	rs969930693	SNP	missense variant	T/I	329	31	deleterious	0.03	323	benign	0.322	21001	likely benign	21
TP53	rs1000256867	SNP	missense variant	E/D	326	111	tolerated	0.11	423	benign	0.422	21001	likely benign	21
TP53	rs1000256867	SNP	missense variant	E/D	326	111	tolerated	0.11	423	benign	0.422	21001	likely benign	21
TP53	rs121912659	SNP	missense variant	G/V	325	171	tolerated	0.17	256	benign	0.255	18001	likely benign	18
TP53	rs121912659	SNP	missense variant	G/E	325	1001	tolerated	1	4	benign	0.003	8001	likely benign	8
TP53	rs863224500	SNP	missense variant	G/R	325	331	tolerated	0.33	19	benign	0.018	23001	likely benign	23
TP53	rs1177881399	SNP	missense variant	D/G	324	1	deleterious	0	891	possibly damaging	0.89	32001	likely deleterious	32
TP53	rs1064794810	SNP	missense variant	D/Y	324	1	deleterious	0	986	probably damaging	0.985	26001	likely benign	26
TP53	rs1064794810	SNP	missense variant	D/H	324	1	deleterious	0	971	probably damaging	0.97	25001	likely benign	25
TP53	rs1064794810	SNP	missense variant	D/N	324	51	tolerated	0.05	300	benign	0.299	23001	likely benign	23
TP53	rs1432281680	SNP	missense variant	L/V	323	511	tolerated	0.51	50	benign	0.049	12001	likely benign	12
TP53	rs863224687	SNP	missense variant	P/S	322	351	tolerated	0.35	38	benign	0.037	17001	likely benign	17
TP53	rs863224687	SNP	missense variant	P/T	322	81	tolerated	0.08	30	benign	0.029	17001	likely benign	17

Figure 7. Representative image of the Mutations tab with various headers representing each feature.

mechanisms in BC development. This also offers baseline data to design improved therapeutic strategies in BC. Herein, we listed all the possible SNPs in a single tab with a large dataset, including each of the TF. The non-synonymous mutations of each TF were retrieved from dbSNP database and processed using SIFT, PolyPhen and CADD servers. These servers help to predict and classify the mutation class, i.e. deleterious, tolerated, benign, possibly damaged or likely benign based on the specific scores. A total of 163 datasets were processed, which contain 142 435 mutations in total and grouped based on the specific output. This mutation-specific information about

each TF is provided to users for retrieval. The mutations tab is featured with different kinds of information including gene name, variant ID, mutation class, consequence type, AA name, AA co-ordinates, SIFT sort, SIFT class, SIFT score, PolyPhen sort, PolyPhen class, PolyPhen score, CADD sort, CADD class and CADD score. The specific mutation tagged with specific variant ID is classified as deleterious or tolerated by the SIFT server based on the SIFT score, benign or probably damaging based on the PolyPhen score by the PolyPhen server and likely benign or likely deleterious by the CADD server based on the CADD score. The implemented search module retrieves mutation information based on gene name or variant ID. This module may help the researchers to directly access the impact of a specific mutation in these TFs. The image tutorial of the Mutations tab is represented in Figure 7.

Drugs tab

Therapeutic development of potential drugs or vaccine candidates is important for disease curtailment. The development of drugs against cancer disease is hot research and needs further investigation. Drug repositioning may be of great interest in research to discover novel drug–target interactions. For this purpose, knowledge about previous drugs and their mode of action is important. Hence, to provide information about the available drugs for these TFs' drug target, the drugs tab was integrated as a part of this database. In this tab, information regarding the target name, DrugBank ID, drug group, pharmacological action and molecular action are provided. The sheet contained information about the DrugBank accession ID of each drug, the drug name and their respective targets. Further, these drugs are divided as either experimental, investigational or approved. Furthermore, information about the pharmacological action, whether the mechanism is known or unknown, are also provided. The classification of each drug as an inhibitor, downregulator, antagonist, agonist, chaperon, binder, inducer, substrate or antibody is also provided in the Actions heading. This information about the drugs of each TF may help the user to retrieve it easily for therapeutic repurposing. The image tutorial of the Drugs tab is represented in Figure 8.

Downloads tab

The Downloads tab enlists all the data as downloadable features. All the data included in the database could be downloaded as a separate file. The image tutorial of the Downloads tab is represented in Figure 9.

Conclusion

In conclusion, the BC-TFdb provides a comprehensive platform for research, to easily access and retrieve information related to TF drivers in BC. The platform provides useful information including gene and protein sequences, protein 3D structures, BC pathway information, survival information, DNA-binding residues, missense mutations and their impact and therapeutics such as the available drugs used against the specific TF targets. This platform may help the researcher to access BC-specific information for computation and experimental analysis. This platform

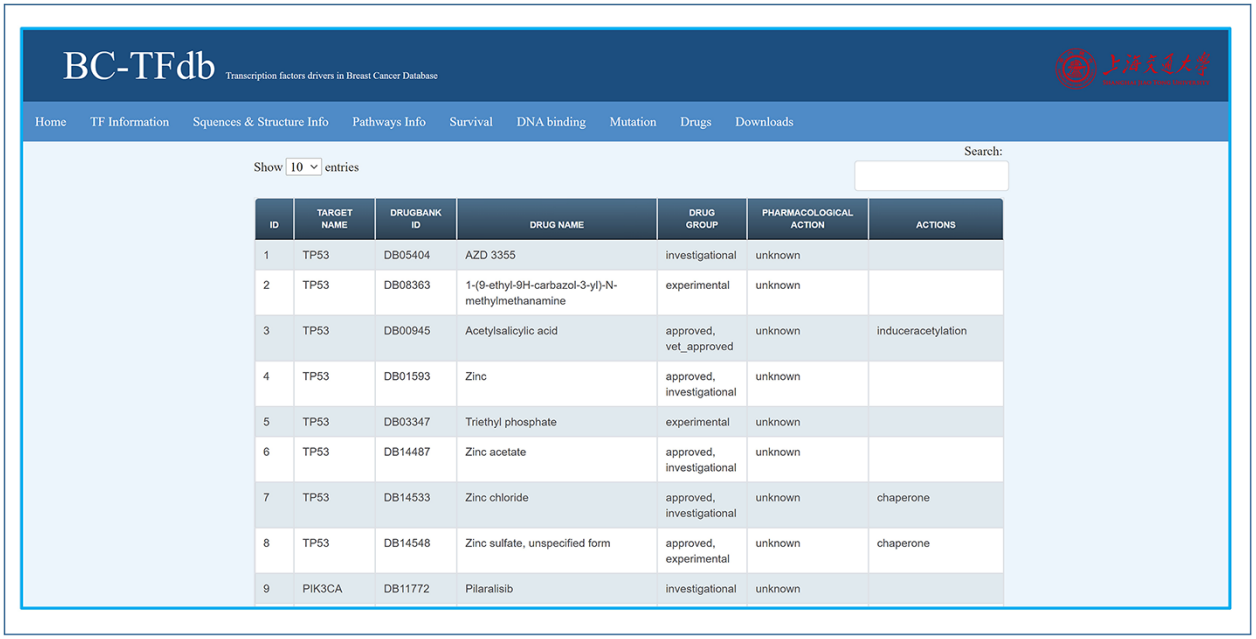


Figure 8. Representative image of the Drugs tab with various headers representing each feature.

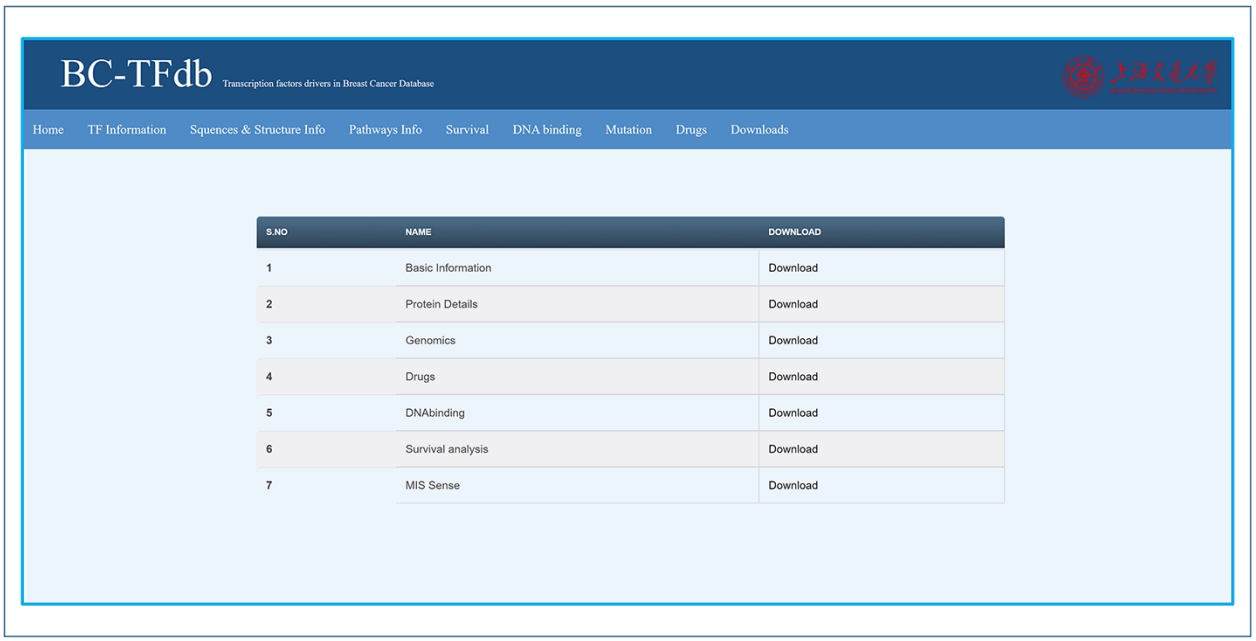


Figure 9. Representative image of the Downloads tab that contains all the files for downloads.

would help to better understand the role of each TF in pathogenesis and would map a better strategy for the management and therapeutic feasibility of BC.

Acknowledgements

The computations were partially performed at the PengCheng Laboratory and the Center for High-Performance Computing, Shanghai Jiao Tong University.

Funding

National Science Foundation of China (32070662, 61832019, 32030063), Key Research Area Grant 2016YFA0501703 of the Ministry of Science and Technology of China, the Science and Technology Commission of Shanghai Municipality (19430750600), as well as SJTU (Shanghai Jiao Tong University) JiRLMDS Joint Research Fund and Joint Research Funds for Medical and Engineering and Scientific Research at Shanghai Jiao Tong University (YG2021ZD02) to D.Q.W. National Natural Science Foundation of China (31801037) and Science Foundation of Army Medical University (2017XQN01, 410310543403, 2019JCZX05) to X.Z.

Conflict of interest. None declared.

References

1. Lee,T.I. and Young,R.A. (2013) Transcriptional regulation and its misregulation in disease. *Cell*, **152**, 1237–1251.
2. Lin,Y., Zhang,Q., Zhang,H.-M. *et al.* (2015) Transcription factor and miRNA co-regulatory network reveals shared and specific regulators in the development of B cell and T cell. *Sci. Rep.*, **5**, 15215.
3. Zhang,Q., Hu,H., Chen,S.-Y. *et al.* (2019) Transcriptome and regulatory network analyses of CD19-CAR-T immunotherapy for B-ALL. *Genom. Proteom. Bioinf.*, **17**, 190–200.
4. Kadonaga,J.T. (2004) Regulation of RNA polymerase II transcription by sequence-specific DNA binding factors. *Cell*, **116**, 247–257.
5. Campbell,L.L. and Polyak,K. (2007) Breast tumor heterogeneity: cancer stem cells or clonal evolution? *Cell Cycle*, **6**, 2332–2338.
6. Shackleton,M., Quintana,E., Fearon,E.R. *et al.* (2009) Heterogeneity in cancer: cancer stem cells versus clonal evolution. *Cell*, **138**, 822–829.
7. Schünemann,H.J., Lerda,D., Quinn,C. *et al.* (2020) Breast cancer screening and diagnosis: a synopsis of the European Breast Guidelines. *Ann. Intern. Med.*, **172**, 46–56.
8. Khan,A., Rehman,Z., Hashmi,H.F. *et al.* (2020) An integrated systems biology and network-based approaches to identify novel biomarkers in breast cancer cell lines using gene expression data. *Interdiscip. Sci.*, **12**, 1–14.
9. Khan,A., Junaid,M., Li,C.-D. *et al.* (2020) Dynamics insights into the gain of flexibility by Helix-12 in ESR1 as a mechanism of resistance to drugs in breast cancer cell lines. *Front. Mol. Biosci.*, **6**, 159.
10. Hughes,T. (2011) Introduction to 'a handbook of transcription factors'. In: *A Handbook of Transcription Factors*. Springer, pp. 1–6.
11. Bhagwat,A. and Vakoc,C. (2015) Targeting transcription factors in cancer. *Trends Cancer*, **1**, 53–65.
12. Dong,M.J., Wang,L.B., Jiang,Z.N. *et al.* (2014) The transcription factor KlF4 as an independent predictive marker for pathologic complete remission in breast cancer neoadjuvant chemotherapy: a case–control study. *Onco Targets Ther.*, **7**, 1963.
13. Span,P.N., Manders,P., Heuvel,J.J. *et al.* (2002) Expression of the transcription factor Ets-1 is an independent prognostic marker for relapse-free survival in breast cancer. *Oncogene*, **21**, 8506–8509.

14. Thompson,P.J., Macfarlan,T.S. and Lorincz,M.C. (2016) Long terminal repeats: from parasitic elements to building blocks of the transcriptional regulatory repertoire. *Mol. Cell*, **62**, 766–776.
15. Coles,C., Condie,A., Chetty,U. *et al.* (1992) p53 mutations in breast cancer. *Cancer Res.*, **52**, 5291–5298.
16. Stemke-Hale,K., Gonzalez-Angulo,A.M., Lluch,A. *et al.* (2008) An integrative genomic and proteomic analysis of PIK3CA, PTEN, and AKT mutations in breast cancer. *Cancer Res.*, **68**, 6084–6091.
17. Nik-Zainal,S., Davies,H., Staaf,J. *et al.* (2016) Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature*, **534**, 47–54.
18. Ma,X., Cheng,J., Zhao,P. *et al.* (2020) DNA methylation profiling to predict recurrence risk in stage I lung adenocarcinoma: development and validation of a nomogram to clinical management. *J. Cell. Mol. Med.*, **24**, 7576–7589.
19. Xiaobin,Z., Chengxiao,L., Jie,Y. *et al.* (2020) Breast cancer stem cells, heterogeneity, targeting therapies and therapeutic implications. *Pharmacol. Res.*, **163**, 105320.
20. Zacksenhaus,E., Liu,J., Jiang,Z. *et al.* (2017) Transcription factors in breast cancer—lessons from recent genomic analyses and therapeutic implications. *Adv. Protein Chem. Struct. Biol.*, **107**, 223–273.
21. Jenuth,J.P. (2000) The NCBI. In: *Bioinformatics Methods and Protocols*. Springer, pp. 301–312.
22. Magrane,M. (2011) UniProt Knowledgebase: a hub of integrated protein data. *Database*, **2011**, 1–13.
23. Rose,P.W., Beran,B., Bi,C. *et al.* (2010) The RCSB Protein Data Bank: redesigned web site and web services. *Nucleic Acids Res.*, **39**, D392–D401.
24. Sherry,S.T., Ward,M.-H., Kholodov,M. *et al.* (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.*, **29**, 308–311.
25. Basyuni,M., Wati,R., Sulistiyono,N. *et al.* (2018) Protein modelling of triterpene synthase genes from mangrove plants using Phyre2 and Swiss-model. In: *Journal of Physics: Conference Series*, Vol. 978. IOP Publishing, Medan, Indonesia. p. 012095.
26. Kanehisa,M., Araki,M., Goto,S. *et al.* (2007) KEGG for linking genomes to life and the environment. *Nucleic Acids Res.*, **36**, D480–D484.
27. Ng,P.C. and Henikoff,S. (2003) SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res.*, **31**, 3812–3814.
28. Hecht,M., Bromberg,Y. and Rost,B. (2015) Better prediction of functional effects for sequence variants. *BMC Genomics*, **16**, S1.
29. Choi,Y. and Chan,A.P. (2015) PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels. *Bioinformatics*, **31**, 2745–2747.
30. Wang,Q., Mehmood,A., Wang,H. *et al.* (2019) Computational screening and analysis of lung cancer related non-synonymous single nucleotide polymorphisms on the human Kirsten rat sarcoma gene. *Molecules*, **24**, 1951.
31. Hou,G.-X., Liu,P., Yang,J. *et al.* (2017) Mining expression and prognosis of topoisomerase isoforms in non-small-cell lung cancer by using Oncomine and Kaplan–Meier plotter. *PLoS One*, **12**, e0174515.
32. Hwang,S., Gou,Z. and Kuznetsov,I.B. (2007) DP-Bind: a web server for sequence-based prediction of DNA-binding residues in DNA-binding proteins. *Bioinformatics*, **23**, 634–636.
33. Kim,S., Chen,J., Cheng,T. *et al.* (2019) PubChem 2019 update: improved access to chemical data. *Nucleic Acids Res.*, **47**, D1102–D1109.
34. Wishart,D.S., Feunang,Y.D., Guo,A.C. *et al.* (2018) DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res.*, **46**, D1074–D1082.
35. Krogh,J.W. (2020) MySQL Workbench. In: *MySQL 8 Query Performance Tuning*. Springer, pp. 199–226.
36. Saroni,M. and Mulyanti,B. (2020) Hypertext preprocessor framework in the development of web applications. In: *IOP Conference Series: Materials Science and Engineering*, Vol. 830. IOP Publishing, Changsha, China. p. 022096.
37. Beasley,R.E. (2020) Ajax Programming. In: *Essential ASP.NET Web Forms Development*. Springer, pp. 499–532.
38. Qian,N., Wang,J., Mueller,F. *et al.* (2020) HTML: a parametric hand texture model for 3D hand reconstruction and personalization. In: Vedaldi, A., Bischof, H., Brox, Th., Frahm, J.-M. (Eds). 16th edn. *European Conference on Computer Vision*. Springer, Glasgow, UK. pp. 54–71.
39. Zhanikeev,M. (2020) A jQuery-like platform for standardized dataset processing logic. *International Journal of Innovation in Management*, **8**, 43–46.