SCIENTIFIC REPORTS

OPEN

Received: 29 November 2016
Accepted: 3 April 2017
Published online: 04 May 2017

CancerPDF: A repository of cancer-associated peptidome found in human biofluids

Sherry Bhalla[1], Ruchi Verma[1], Harpreet Kaur[1], Rajesh Kumar[1], Salman Sadullah Usmani[1], Suresh Sharma[2] & Gajendra P. S. Raghava[1]

CancerPDF (Cancer Peptidome Database of bioFluids) is a comprehensive database of endogenous peptides detected in the human biofluids. The peptidome patterns reflect the synthesis, processing and degradation of proteins in the tissue environment and therefore can act as a gold mine to probe the peptide-based cancer biomarkers. Although an extensive data on cancer peptidome has been generated in the recent years, lack of a comprehensive resource restrains the facility to query the growing community knowledge. We have developed the cancer peptidome resource named CancerPDF, to collect and compile all the endogenous peptides isolated from human biofluids in various cancer profiling studies. CancerPDF has 14,367 entries with 9,692 unique peptide sequences corresponding to 2,230 unique precursor proteins from 56 high-throughput studies for ~27 cancer conditions. We have provided an interactive interface to query the endogenous peptides along with the primary information such as m/z, precursor protein, the type of cancer and its regulation status in cancer. To add-on, many web-based tools have been incorporated, which comprise of search, browse and similarity identification modules. We consider that the CancerPDF will be an invaluable resource to unwind the potential of peptidome-based cancer biomarkers. The CancerPDF is available at the web address http://crdd.osdd.net/raghava/cancerpdf/.

Cancer is considered as the major public health concern worldwide and the second most health hazardous disease, causing deaths in the United States. Globally 14 million new cases and 8.2 million cancer-related deaths have been reported in 2012[1]. In 2017, there is an approximation of 63,990 new cases and 14,400 deaths from cancer in the United States[2]. Although the rate of survival has increased over the years, still it is meager. The lack of diagnosis at an early stage is one of the major hurdles in treating the cancer patients[3]. Cancer detection is often skewed due to the lack of accurate and non-invasive markers. Due to advances in genomics and proteomics, the probability to detect the cancer at an early stage has improved using peptide-based biomarkers[4]. In recent years, the peptide-based biomarkers have emerged as diagnostic tools in several foodborne diseases[5], arthritis[6], inflammatory disease[7] as well as cancer[8, 9]. Thus it is imperative to understand the mechanism of action and processing of peptides in mammalian biofluids. Following are the few examples of peptide-based biomarkers; i) Insulin and C-peptide are used in case of diabetes[10], ii) Calcitonin, and collagen fragments in case of osteoporosis[11–13], iii) Pro-gastrin-releasing peptide for small cell lung carcinoma[14], iv) β-amyloid 1–42 for Alzheimer's disease[15] and v) angiotensin II for hypertension[16].

In past, large number of peptide repositories and computational resources have been developed to explore full potential of peptides in medical sciences[17–19]. Pepbank is the generalized database of biologically relevant peptides containing nearly twenty thousand peptides obtained using text mining[20]. Some of the peptide databases like PeptideAtlas[21, 22] and SwePep[23] are specifically derived from mass spectrometry proteomics data. PeptideAtlas is one of the largest repositories of peptides identified from tandem mass spectrometry experiments collected from human, mouse, yeast and several other organisms. Similarly, SwePep database contains approximately four thousand endogenous peptides from different tissues originated from diverse species. Some databases like PeptideDB[24], Endogenous Regulatory OligoPeptide knowledgebase[25, 26] and BIOPEP database[27] are specifically made to store naturally occurring bioactive peptides. Recently databases have been developed for maintaining

[1]Bioinformatics Centre, CSIR-Institute of Microbial Technology, Sector 39A, Chandigarh, 160036, India. [2]Centre for Systems Biology and Bioinformatics, Panjab University, Sector 14, Chandigarh, 160014, India. Sherry Bhalla, Ruchi Verma and Harpreet Kaur contributed equally to this work. Correspondence and requests for materials should be addressed to G.P.S.R. (email: raghava@imtech.res.in)

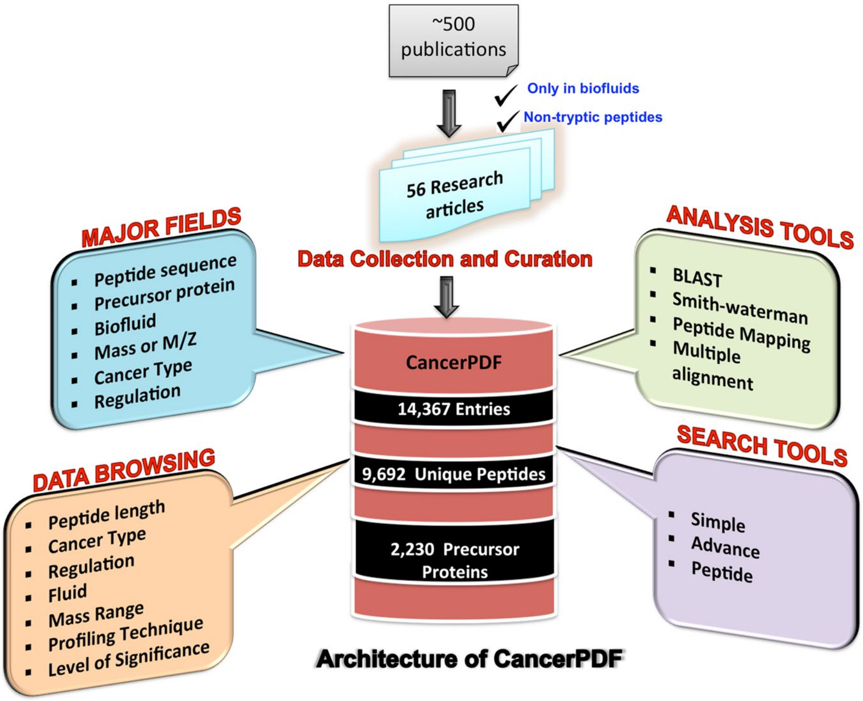

Figure 1. Architecture of CancerPDF database.

peptides important for designing anticancer drugs[28]. The CancerPPD[28] contains 3,491 anticancer peptides and 121 anticancer proteins with diverse origin. Similarly, the TumorHoPe[29] database contains peptides that can recognize tumor tissues and tumor associated microenvironment.

Despite several databases have been developed to maintain different classes of peptides in the past, there is no dedicated repository of peptides (peptidome) released in the tumor microenviroment during cancer progression. Thus, there is a need to compile cancer-associated peptides or cancer-peptidome found in human biofluid[30]. Cancer-peptidome can act as a rich source of peptide biomarkers as it represents the various cellular and enzymatic processes happening in the tumor microenvironment. The peptide patterns generated by peptidomics study can aid in understanding the pathology of the disease[31]. The study of endogenous peptide patterns also hints the alterations in protease activity in cancer microenvironment, which deepens the pathophysiological awareness of the disease[32]. The circulating peptides in cancer patients have shown to classify patient subtypes providing a direct therapeutic approach to those individuals at an earlier stage, which is otherwise not detectable[33, 34]. There have been many high-throughput studies in which peptidome of the various biofluids like plasma, serum, blood, urine and their peptide content in cancer patients have been reported[35–37]. In this light, different groups have collected data regarding plasma proteome and cancer secretome and made attempts to develop resources such as Plasma Proteome Database [10.1093/nar/gkt1251] and Human Cancer Secretome Database [10.1093/database/bav051] compiling this information at the protein level.

To the best of authors' knowledge, no attempts have been made to organize all the endogenous peptides, detected in various biofluids from different human cancers using clinical samples. A repertoire of these peptides will certainly be helpful for the scientific community in studying and discovering new peptide-based cancer biomarkers. In order to facilitate scientific community, we have developed a resource called CancerPDF. This database offers comprehensive information on naturally occurring peptides in the biofluids of cancer patients and their expression status as reported by the original studies. This structured information can be used for identification of cancer biomarkers from proteomics data of biofluids. This database integrates various web-based tools to facilitate users in extracting and analyzing data. In order to provide access from the wide range of devices (like Smartphones, iPads, Tablets), we have developed web interface using responsive web templates.

Results

Database statistics. CancerPDF is a comprehensive resource of naturally occurring peptides found in biofluids using mass spectrometry. We have collected peptides, found only in the human biofluids from 56 studies which comprises of 14,367 entries corresponding to m/z values, out of which 9,692 entries have corresponding peptide sequences identified from 2,230 proteins (Fig. 1). The length of collected peptides in CancerPDF varies from 4 to 113 amino acid residues. Maximum peptides are in the range of 10–40 amino acid residues (Fig. 2A). The m/z values of endogenous peptides mostly varied from 300 to 14,000. Most of the peptides have mass in the range of 300 Da to 6,000 Da (Fig. 2B). The 56 studies encompassed nearly 27 different types of cancer conditions. The primary cancers according to tissues types are Ovary, Bladder, Melanoma, Colorectal and Multiple myeloma (Table 1 and Fig. 2C). Most of the peptides were derived from biofluids like urine, serum, plasma, ascites fluid, saliva and others with records corresponding to 5955, 4539, 2875, 777, 170 and 51 peptides respectively. Maximum studies were related to urine, plasma and serum, as they are most easy to obtain and non-invasive fluids which can be

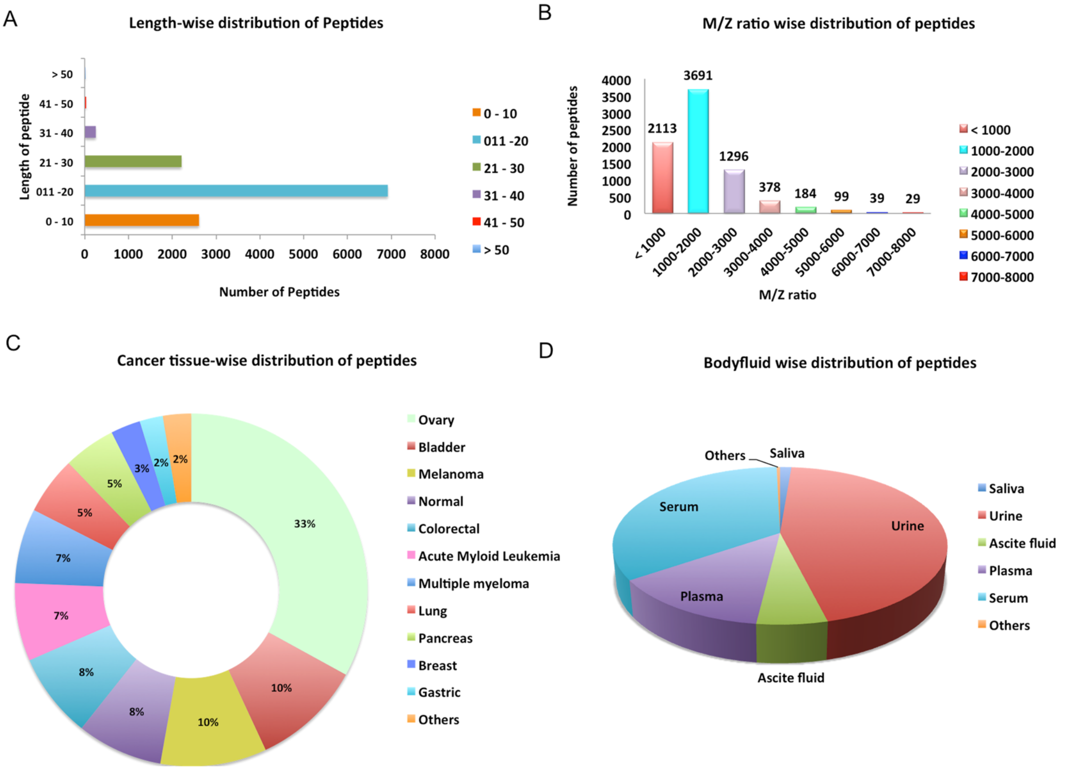

Figure 2. Distribution of peptides according to length (**A**), mass range (**B**), cancer tissue types (**C**) and biofluids (**D**) in CancerPDF database.

Biofluid	Serum	Plasma	Urine	Others	Total
Cancer					
Ovary	67	0	4368	777	5212
Bladder	80	0	1515	0	1595
Melanoma	1539	0	0	0	1539
Colorectal	1186	44	3	0	1233
Multiple myeloma	0	1083	0	0	1083
Lung	836	0	0	0	836
Pancreas	66	690	0	0	756
Breast	419	8	0	5	432
Gastric	335	0	0	1	336
Thyroid	103	0	0	0	103
Renal	36	0	62	0	98
Others	81	9	7	19	116
Total	4748	1834	5955	802	13339

Table 1. Distribution of CancerPDF entries across key cancer types and major body fluids.

used to detect the cancer (Table 1 and Fig. 2D). These peptides are mainly profiled and identified using label-free mass spectrometry techniques such as LC/MS-MS and MALDI-TOF MS-MS. To assess the information about the precursor proteins from which these peptides are derived, we converted all the protein names to the UniProtKB entry names. CancerPDF peptides map to 2,230 unique UniProtKB entry names. The proteins for which the maximum numbers of peptides are found include FIBA_HUMAN, CO3_HUMAN, APOA1_HUMAN, CO1A1_HUMAN and A4_HUMAN (Table 2). Eight out of the top ten proteins with the highest number of identified peptides of CancerPDF are found to be differentially expressed in dbDEPC 2.0[38], which is a database of differentially expressed proteins in cancer.

Oxidation and hydroxylation are the most commonly occurring modifications in peptides, *i.e.* in 404 and 198 peptides, respectively. Another important aspect of these peptides is their differential regulation in various conditions like cancer versus normal. Wherever available, we have collected the information whether the peptides were differentially expressed, uniquely expressed, up-regulated and down-regulated in different conditions as

UniprotKB entry name	Number of Unique peptides	Number of Studies	Number of Cancer conditions
FIBA_HUMAN	727	25	21
CO3_HUMAN	296	16	15
APOA1_HUMAN	266	14	13
CO1A1_HUMAN	232	4	3
A4_HUMAN	223	12	12
A1AT_HUMAN	204	8	7
H4_HUMAN	200	20	16
APOA4_HUMAN	199	9	9
ITIH4_HUMAN	182	18	14
ALBU_HUMAN	168	11	11

Table 2. Top ten proteins with maximum numbers of reported peptides in CancerPDF.

reported in the corresponding studies. In this database, the peptides are reported to be differentially expressed in cancer versus healthy conditions, based on the level of significance ($p\text{-value} < 0.05$) reported in original study. CancerPDF comprises of 2,379 entries of differentially expressed peptides among diverse groups. Further there are 464 up-regulated, 355 down-regulated and nearly 5,152 uniquely expressed peptide peaks in various cancers. We have also specified the classification sensitivity, specificity and accuracy of the peptides biomarkers as reported in the respective studies (wherever possible) to provide an estimate of biomarker peptide efficiency.

Implementation of web tools. To enable convenient data searching, various tools such as retrieval, browsing and analysis were integrated with CancerPDF.

Search tools. We have implemented three different modules namely 'Simple search', 'Peptide Search' and 'Advance search' under the search option to provide a facility for the adequate data retrieval.

Simple search. This tool represents key data retrieval module from the CancerPDF. The keyword search can be executed by a user on the major fields of the database such as PubMed ID, Biofluids, Protein Name, Cancer Type, Regulation and Validation etc. Moreover, this module also allows the users to select various fields to be displayed for the result.

Peptide search. This tool offers a platform for searching a given peptide sequence against all peptide sequences available in CancerPDF. It searches for the exact match as well as substring matches in the database. Exact search option retrieves those peptides from the database, which have an identical amino acid sequence with the query peptide. While substring search option retrieves those peptides that contain the query peptide.

Advance search. This module assists the user to perform multiple structured query system options for the retrieval of the required information from the CancerPDF. By default, it performs four queries simultaneously, but a user can choose desired keyword search from any selected field. Besides this, advance search offers the user to apply standard logical operators (e.g. =, >, < and LIKE). Moreover, this module permits the user to integrate the output of different queries by utilizing operators like 'AND and OR'. Additionally, the user can also add or remove the queries to be implemented.

Browse tools. In CancerPDF, we have implemented browsing facility, which helps the user for convenient data navigation within the database in an orderly manner. In this module, a user can retrieve information on peptides by browsing nine different categories (i) Cancer Type, (ii) Fluid, (iii) Regulation, (iv) Precursor Protein, (v) Profiling Technique, (vi) Mass Range, (vii) Level of significance (p-value), (viii) Peptide Length and (ix) PubMed ID.

The 'Cancer Type' field facilitates the user to extract the information on peptides obtained from specific cancer conditions such as Lung cancer, Breast cancer, Prostate cancer etc. From the 'Fluid' category, the user is allowed to retrieve detailed information on the peptides isolated from a particular type of biofluid e.g. serum, plasma, urine and saliva. The 'Regulation' field offers the user to fetch the information on peptides that are up-regulated, down-regulated, differentially expressed in cancer condition as compared to healthy and peptides that are uniquely expressed in a specific type of cancer. In addition, by 'Precursor Protein' category, user can withdraw information on those peptides that are derived from a specific precursor protein such as Fibrinogen-alpha chain, Fibrinogen-beta chain and Complement component C3f etc. The 'Profiling Technique' option permits the user to extract information regarding the peptides that are profiled using different techniques such as MALDI-TOF, LC-MS etc. Furthermore, a user can also extract the information of peptides on the basis of their length by browsing 'Mass Range', 'Level of significance (p-value)', 'Peptide Length' and 'PubMed ID'.

Similarity. This module facilitates the user to perform various analyses such as sequence similarity, mapping and multiple sequence alignment by implementing different web-based tools i.e. Basic Local Alignment Search Tool (BLAST), Smith-Waterman, Multiple Sequence Alignment in CancerPDF database.

BLAST Search. This tool offers a user to execute a similarity-based search against CancerPDF database. Peptide sequences should be submitted in FASTA format and the user can choose different parameters such as weight matrix and an expectation value for the execution of BLAST search[39].

Smith-Waterman Search. This algorithm executes similarity search against small peptides more efficiently using Smith-Waterman algorithm[40]. This module permits the user to search peptides in CancerPDF database similar to their query peptides. In this option, a user can submit simultaneously multiple peptide sequences in FASTA format.

Multiple Sequence Alignment (MSA). This module offers the user to align their peptide sequences using ClustalW[41] sequences along the peptides of CancerPDF Database. A user can perform batch submission in FASTA format in provided input box to get aligned sequences using MSA viewer[42].

Peptide Mapping. This tool permits a user to map CancerPDF peptides over their peptide sequences. Under this module, the user can perform mapping using two options i.e. Sub search and Super search. In Sub search, query peptide is mapped across all the peptides in the CancerPDF, while Super search allows mapping of protein sequence against CancerPDF. The Super search module is useful to identify the local region of the query protein that is identical to peptides of CancerPDF.

Comparison with other peptide and protein databases. CancerPDF database consists of endogenous peptides that are found in the biofluids of cancer patients. To understand the biological importance of these peptides, we compared the peptides using sequence-based similarity in CancerPDF with already existing peptide resources such as PeptideAtlas and immune epitope database and analysis resource (IEDB)[43]. We found numerous overlapping and exclusive peptides in CancerPDF as compared to these two resources (Supplementary Figure S1). Mapping peptides in CancerPDF with PeptideAtlas human build resulted in 2,007 common peptides. On comparing the CancerPDF with IEDB, 1,526 exact matches were found. Out of these, 1,301 were found to be MHC-I restricted peptides. This indicates the activation of the cell-mediated immune system during cancer progression; mediated via MHC-I restricted peptides. In literature, it is well known that cell-mediated immunity is triggered in the body during tumorigenesis, but becomes ineffective due to local suppressive factors at tumor sites[44–46]. This analysis shows that these peptides can be further explored for designing therapeutic vaccines against cancer based on MHC-I restricted peptides, due to their stability under cancerous conditions[47].

Moreover, to understand the significance of proteins in our database, we have compared the precursor proteins of CancerPDF peptides with the database of differentially expressed proteins in cancers named dbDEPC 2.0[38] and obtained 232 common UniProtKB entry names of proteins. This type of analysis indicated that the differentially expressed endogenous peptides reflect differentially expressed precursor proteins in cancer patients.

Discussion

Peptidomics is an emerging field that deals with the comprehensive qualitative and quantitative analysis of peptides in biological samples[9]. During protein processing and degradation of other biological macromolecules, peptides are derived either from precursor protein or as degradation products. Therefore, subjecting to the physiological state of an organism, the amount of the peptide repertoire changes within body circulation. The pathological or diseased state has the direct effect on these peptide repertoires[48]. Detecting biomarkers in biofluids is one of the most extensive research interests in this era as it is the most non-invasive approach to uncover biomarker for various diseases[49]. The naturally occurring peptide patterns can be exploited to detect variations at the proteomics level of the tumor microenvironment[50]. The CancerPDF database provides the collection of endogenous peptides in the human biofluids and their precursor proteins that are found in the cancer peptidome profiling studies. As a comprehensive resource containing 14,367 entries, CancerPDF can aid in defining candidate peptide biomarkers derived from the biofluids in cancer. This database also stores the peptides that are differentially regulated and uniquely found in different types of cancer. CancerPDF can be a very important source to mine the peptides that are differentially regulated in specific type of cancer in different population cohorts and peptides that are differentially regulated across different types of cancer. Further analysis of a particular protein with its associated peptides in cancer will shed light on activation and deactivation of various proteolytic events specific to cancer. We foresee that CancerPDF will act as preliminary effort that will help in analyzing cancer peptidome associations and peptide-based cancer biomarker discovery.

Utility of database

In the last decade several databases have been developed that maintain different type of information related to peptides and proteins. Thus it is essential to rationalize the need of another peptide database or the unique features of the CancerPDF. Some of the potential applications of the CancerPDF include.

Screening of cancer biomarker. The CancerPDF includes the peptides and their precursor proteins that are differentially regulated in various cancer conditions. The user can easily identify number of differentially regulated peptides found in a particular type of cancer. The presence and absence of the differentially expressed peptides can be used as features for developing prediction models for discriminating cancer and healthy individuals. Thus CancerPDF is an important resource for developing biomarkers for the different types of cancer. These peptides are founds in bodyfluids that make them potential non-invasive biomarkers for detecting cancer.

This database can help in understanding the change in peptide content during developement of cancer (e.g., breast cancer). In order to demonstrate its application, we browsed the entries of the breast cancer. We obtained total 432 entries with 177 unique peptides that include 120 up-regulated, 28 down-regulated

Cancer→	Prostate Cancer	Bladder Cancer	Breast Cancer	NSCLC	Lung adeno-carcinoma	Renal Cell carcinoma	Colorectal carcinoma	Metastatic thyroid carcinomas	Ovarian Cancer	ESCC	Cervical Cancer
Sequence											
ADSGEGDFLAEGGGVR	1	3	1	1	1	—	—	—	1	—	—
SGEGDFLAEGGGVR	1	2	1	1	1	1	1	1	1	—	—
RPPGFSPFR	1	2	3	—	—	—	1	—	—	—	—
DSGEGDFLAEGGGVR	1	2	1	1	1	1	1	1	—	—	—
SKITHRIHWESASLL	1	1	2	1	—	—	1	1	—	—	—
MNFRPGVLSSRQLGLPGPPDVPDHAAYHPF	1	1	5	—	—	—	—	—	—	1	—
SSKITHRIHWESASLL	1	1	2	1	—	—	—	1	—	—	—
RPPGFSPF	1	1	2	1	—	—	1	—	—	—	1
KITHRIHWESASLL	1	1	2	1	—	—	—	1	—	—	—
GEGDFLAEGGGVR	1	1	1	1	1	1	1	1	—	—	—
DEAGSEADHEGTHSTKRGHAKSRPV	1	3	4	—	—	—	—	—	—	—	—
SSSYSKQFTSSTSYNRGDSTFESKSYKM	1	1	2	—	1	—	—	1	—	—	1
NGFKSHALQLNNRQIR	1	1	2	—	—	—	1	—	—	—	1
DFLAEGGGVR	1	1	1	—	1	—	1	1	—	—	1
THRIHWESASLL	1	1	2	—	—	—	—	1	—	—	—

Table 3. Top fifteen unique peptides associated with different cancers. Each number in the cell represents the number of studies associated with each cancer.

and 25 differentially expressed peptides (p-value < 0.05). It was observed that peptide sequence "MNFRPGVLSSRQLGLPGPPDVPDHAAYHPF", has been found to be up-regulated in three different studies of breast cancer. This type of peptide is important while defining candidate peptide biomarkers as it has been found up-regulated in three independent population cohorts. So out of all reported peptides user can get these types of lead peptides to further confirm for biomarker potential.

Peptide Library. A user can use the peptides in CancerPDF as the peptide library to directly search raw mass spectrometry cancer data to find the already known endogenous peptides in a particular sample. This will facilitate the researcher in identification of differentially regulated peptides in their sample that have been already annotated in previous studies.

Pan-cancer analysis. CancerPDF offers the opportunity to search for those peptides that are differentially regulated across multiple types of cancers and also for those peptides that are differentially regulated in a specific cancer. One of the peptide sequences (SGEGDFLAEGGGVR) was found in 10 different studies and differentially regulated in 9 types of cancer (Table 3). These types of inferences can be crucial for further mining of peptide biomarkers for cancer.

In summary, CancerPDF is an invaluable resource to the scientific community working in the area of peptide-based cancer diagnostics.

Methods

Data collection. We queried PubMed to obtain the research articles with the keywords "cancer [Title/Abstract] AND peptidome [Title/Abstract]" and "cancer [Title/Abstract]) AND endogenous [Title/Abstract] AND peptides [Title/Abstract]" and collected around 500 publications till September 2016. All research articles were curated manually to understand the type of information available in these articles. After reading all articles carefully, we kept articles for further processing that have information relevant to naturally occurring peptides extracted from body fluids. We excluded all those articles, for which peptides/peptidome were derived either using tryptic digestion, or from cell lines and tissues. We have also included publications that include peptidome of biofluids of normal individuals.

We manually retrieved information from selected articles regarding the sequence of peptides, precursor protein, their m/z value, mass (in Daltons or H^+), charge, modification, profiling techniques, peptide identification technique, quantification techniques, their regulation, type of cancer, fluid sample from which peptides were extracted, and validation etc.

Architecture and interface of database. CancerPDF is assembled employing Apache HTTP Server on Red Hat Linux system. A responsive web template is used as the web interface for the front end of this database. Thus web interface is compatible to the wide range of modern devices that includes Mobile, Tablet, Ipad, iMac and Desktop. The front end of the database is developed using HTML5, CSS3, PHP (version 5.2.14) and JavaScript (version 1.7). To manage the data efficiently, we used an object-relational database management system (RDBMS) MySQL at the back end. CancerPDF has numerous web-based tools to compile, explore and retrieve the information from the database.

Organization of database. In CancerPDF, data is categorized into primary and secondary information. The primary information, procured from the research articles was arranged into defined categories which include (i) Peptide: its sequence, length and modification; (ii) Precursor protein: Protein name, as given in research article and UniProtKB entry name retrieved using *bioDBnet* tool[51] and DAVID[52]; (iii) Physical properties of Peptide: m/z ratio, Mass (H^+), Mass (in Daltons) and charge; (iv) Cancer aspects: Type of cancer, Number of cancer patients and Regulation status of peptide in cancer condition; (v) Biofluid from which peptide was isolated; (vi) Statistics of peptide identification: p-value and false discovery rate (FDR); (vii) Performance Measures: validation, sensitivity, specificity and accuracy; and (viii) Pubmed ID of research article from which information was extracted. In addition to primary information, in the secondary information category, each peptide is linked to IEDB and Peptide Atlas database wherever available.

Availability

CancerPDF can be accessed freely at http://crdd.osdd.net/raghava/cancerpdf/.

References

1. Torre, L. A. Bray, Freddie, Siegel, Rebecca L., Ferlay, Jacques, Lortet-Tieulent, Joannie, Jemal, Ahmedin. Global cancer statistics, 2012. *CA: A Cancer Journal for Clinicians* **65**, 87–108 (2012).
2. Rebecca, L., Siegel, K. D. M. & Ahmedin Jemal, D. V. M. Cancer statistics, 2017. *CA: A Cancer Journal for Clinicians* **67**, 7–30 (2017).
3. Virnig, B. A., Baxter, N. N., Habermann, E. B., Feldman, R. D. & Bradley, C. J. A matter of race: early-versus late-stage cancer diagnosis. *Health Aff (Millwood)* **28**, 160–168 (2009).
4. Omenn, G. S. Strategies for Genomic and Proteomic Profiling of Cancers. *Stat Biosci* **8**, 1–7 (2016).
5. Singhal, N., Kumar, M., Kanaujia, P. K. & Virdi, J. S. MALDI-TOF mass spectrometry: an emerging technology for microbial identification and diagnosis. *Front Microbiol* **6**, 791 (2015).
6. Stalmach, A. *et al.* Identification of urinary peptide biomarkers associated with rheumatoid arthritis. *PLoS One* **9**, e104625 (2014).
7. Bennike, T., Birkelund, S., Stensballe, A. & Andersen, V. Biomarkers in inflammatory bowel diseases: current status and proteomics identification strategies. *World J Gastroenterol* **20**, 3231–3244 (2014).
8. Diamandis, E. P. Peptidomics for cancer diagnosis: present and future. *J Proteome Res* **5**, 2079–2082 (2006).
9. Schulte, I., Tammen, H., Selle, H. & Schulz-Knappe, P. Peptides in body fluids and tissues as markers of disease. *Expert Rev Mol Diagn* **5**, 145–157 (2005).
10. Jones, A. G. & Hattersley, A. T. The clinical utility of C-peptide measurement in the care of patients with diabetes. *Diabet Med* **30**, 803–817 (2013).
11. Romero Barco, C. M., Manrique Arija, S. & Rodriguez Perez, M. Biochemical markers in osteoporosis: usefulness in clinical practice. *Reumatol Clin* **8**, 149–152 (2012).
12. Kraenzlin, M. E. & Meier, C. Parathyroid hormone analogues in the treatment of osteoporosis. *Nat Rev Endocrinol* **7**, 647–656 (2011).
13. Hodsman, A. B., Fraher, L. J., Ostbye, T., Adachi, J. D. & Steer, B. M. An evaluation of several biochemical markers for bone formation and resorption in a protocol utilizing cyclical parathyroid hormone and calcitonin therapy for osteoporosis. *J Clin Invest* **91**, 1138–1148 (1993).
14. Oremek, G. M. & Sapoutzis, N. Pro-gastrin-releasing peptide (Pro-GRP), a tumor marker for small cell lung cancer. *Anticancer Res* **23**, 895–898 (2003).
15. Tapiola, T. *et al.* Cerebrospinal fluid {beta}-amyloid 42 and tau proteins as biomarkers of Alzheimer-type pathologic changes in the brain. *Arch Neurol* **66**, 382–389 (2009).
16. Xu, Z., Xu, B. & Xu, C. Urinary angiotensinogen as a potential biomarker of intrarenal renin-angiotensin system activity in Chinese chronic kidney disease patients. *Ir J Med Sci* **184**, 297–304 (2015).
17. Singh, S. *et al.* SATPdb: a database of structurally annotated therapeutic peptides. *Nucleic Acids Res* **44**, D1119–1126 (2016).
18. Nagpal, G. *et al.* Computer-aided designing of immunosuppressive peptides based on IL-10 inducing potential. *Sci Rep* **7**, 42851 (2017).
19. Mathur, D. *et al.* PEPlife: A Repository of the Half-life of Peptides. *Sci Rep* **6**, 36617 (2016).
20. Shtatland, T., Guettler, D., Kossodo, M., Pivovarov, M. & Weissleder, R. PepBank–a database of peptides based on sequence text mining and public peptide data sources. *BMC Bioinformatics* **8**, 280 (2007).
21. Farrah, T. *et al.* The state of the human proteome in 2012 as viewed through PeptideAtlas. *J Proteome Res* **12**, 162–171 (2013).
22. Desiere, F. *et al.* The PeptideAtlas project. *Nucleic Acids Res* **34**, D655–658 (2006).
23. Falth, M. *et al.* SwePep, a database designed for endogenous peptides and mass spectrometry. *Mol Cell Proteomics* **5**, 998–1005 (2006).
24. Dziuba, J., Minkiewicz, P., Nalecz, D. & Iwaniak, A. Database of biologically active peptide sequences. *Nahrung* **43**, 190–195 (1999).
25. Zamyatnin, A. A., Borchikov, A. S., Vladimirov, M. G. & Voronina, O. L. The EROP-Moscow oligopeptide database. *Nucleic Acids Res* **34**, D261–266 (2006).
26. Zamyatnin, A. A. EROP-Moscow: specialized data bank for endogenous regulatory oligopeptides. *Protein Seq Data Anal* **4**, 49–52 (1991).
27. Liu, F., Baggerman, G., Schoofs, L. & Wets, G. The construction of a bioactive peptide database in Metazoa. *J Proteome Res* **7**, 4119–4131 (2008).
28. Tyagi, A. *et al.* CancerPPD: a database of anticancer peptides and proteins. *Nucleic Acids Res* **43**, D837–843 (2015).
29. Kapoor, P. *et al.* TumorHoPe: a database of tumor homing peptides. *PLoS One* **7**, e35187 (2012).
30. Lai, Z. W., Petrera, A. & Schilling, O. The emerging role of the peptidome in biomarker discovery and degradome profiling. *Biol Chem* **396**, 185–192 (2015).
31. Di Meo, A., Pasic, M. D. & Yousef, G. M. Proteomics and peptidomics: moving toward precision medicine in urological malignancies. *Oncotarget* **7**, 52460–52474 (2016).
32. Diamandis, E. P. Point: Proteomic patterns in biological fluids: do they represent the future of cancer diagnostics? *Clin Chem* **49**, 1272–1275 (2003).
33. Bay-Jensen, A. C., Henrotin, Y., Karsdal, M. & Mobasheri, A. The Need for Predictive, Prognostic, Objective and Complementary Blood-Based Biomarkers in Osteoarthritis (OA). *EBioMedicine* **7**, 4–6 (2016).
34. Doble, N. & Baron, J. H. Anticoagulation control with warfarin by junior hospital doctors. *J R Soc Med* **80**, 627 (1987).
35. Fan, N. J., Gao, C. F., Zhao, G., Wang, X. L. & Liu, Q. Y. Serum peptidome patterns of breast cancer based on magnetic bead separation and mass spectrometry analysis. *Diagn Pathol* **7**, 45 (2012).
36. Bedin, C. *et al.* Alterations of the Plasma Peptidome Profiling in Colorectal Cancer Progression. *J Cell Physiol* **231**, 915–925 (2016).
37. Smith, C. R. *et al.* Deciphering the peptidome of urine from ovarian cancer patients and healthy controls. *Clin Proteomics* **11**, 23 (2014).
38. He, Y. *et al.* dbDEPC 2.0: updated database of differentially expressed proteins in human cancers. *Nucleic Acids Res* **40**, D964–971 (2012).
39. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J Mol Biol* **215**, 403–410 (1990).
40. Pearson, W. R. Flexible sequence similarity searching with the FASTA3 program package. *Methods Mol Biol* **132**, 185–219 (2000).

41. Sievers, F. *et al*. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol* **7**, 539 (2011).
42. Yachdav, G. *et al*. MSAViewer: interactive JavaScript visualization of multiple sequence alignments. *Bioinformatics* **32**, 3501–3503 (2016).
43. Vita, R. *et al*. The immune epitope database (IEDB) 3.0. *Nucleic Acids Res* **43**, D405–412 (2015).
44. Rosenberg, S. A. Progress in human tumour immunology and immunotherapy. *Nature* **411**, 380–384 (2001).
45. Boon, T., Coulie, P. G., Van den Eynde, B. J. & van der Bruggen, P. Human T cell responses against melanoma. *Annu Rev Immunol* **24**, 175–208 (2006).
46. Aptsiauri, N. *et al*. MHC class I antigens and immune surveillance in transformed cells. *Int Rev Cytol* **256**, 139–189 (2007).
47. Comber, J. D. & Philip, R. MHC class I antigen presentation and implications for developing a new generation of therapeutic vaccines. *Ther Adv Vaccines* **2**, 77–89 (2014).
48. Valant, P. A., Adjei, P. N. & Haynes, D. H. Rapid Ca^{2+} extrusion via the Na^+/Ca^{2+} exchanger of the human platelet. *J Membr Biol* **130**, 63–82 (1992).
49. Good, D. M. *et al*. Naturally occurring human urinary peptides for use in diagnosis of chronic kidney disease. *Mol Cell Proteomics* **9**, 2424–2437 (2010).
50. Petricoin, E. F., Belluco, C., Araujo, R. P. & Liotta, L. A. The blood peptidome: a higher dimension of information content for cancer biomarker discovery. *Nat Rev Cancer* **6**, 961–967 (2006).
51. Karbhal, R., Sawant, S. & Kulkarni-Kale, U. BioDB extractor: customized data extraction system for commonly used bioinformatics databases. *BioData Min* **8**, 31 (2015).
52. Huang da, W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* **4**, 44–57 (2009).

Acknowledgements

Authors are grateful to Council of Scientific and Industrial Research (CSIR), Department of Biotechnology (DBT), Department of Science and Technology (DST), Indian Council of Medical Research (ICMR) and Government of India for financial support and fellowships. Funding for open access charge: Council of Scientific and Industrial Research (CSIR) [fellowships and project: Open Source Drug Discovery and GENESIS BSC0121]; Indian Council of Medical Research (ICMR).

Author Contributions

S.B., H.K., R.V. and R.K. gathered and compiled the data. S.B., H.K., R.V., R.K., and S.U. developed the web interface. S.B., S.S., and G.P.S.R. analyzed the data, S.B., H.K., R.V., R.K., S.U. and G.P.S.R. prepared the manuscript. G.P.S.R. envisioned the idea and managed the project.

Additional Information

Supplementary information accompanies this paper at doi:10.1038/s41598-017-01633-3

Competing Interests: The authors declare that they have no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.