Briefings in Functional Genomics, 20(4), 2021, 213–222

https://doi.org/10.1093/bfgp/elab021
Advance Access Publication Date: 1 April 2021
Review Paper

Computational resources for identification of cancer biomarkers from omics data

Harpreet Kaur[†], Rajesh Kumar[†], Anjali Lathwal[†] and Gajendra P.S. Raghava

Corresponding author: Gajendra P.S. Raghava, Department of Computational Biology, Indraprastha Institute of Information Technology, Okhla Industrial Estate, Phase III, New Delhi 110020, India. Tel.: +91 011 26907444; E-mail address: raghava@iiitd.ac.in
[†]These authors have contributed equally.

Abstract

Cancer is one of the most prevailing, deadly and challenging diseases worldwide. The advancement in technology led to the generation of different types of omics data at each genome level that may potentially improve the current status of cancer patients. These data have tremendous applications in managing cancer effectively with improved outcome in patients. This review summarizes the various computational resources and tools housing several types of omics data related to cancer. Major categorization of resources includes—cancer-associated multiomics data repositories, visualization/analysis tools for omics data, machine learning-based diagnostic, prognostic, and predictive biomarker tools, and data analysis algorithms employing the multiomics data. The review primarily focuses on providing comprehensive information on the open-source multiomics tools and data repositories, owing to their broader applicability, economic-benefit and usability. Sections including the comparative analysis, tools applicability and possible future directions have also been discussed in detail. We hope that this information will significantly benefit the researchers and clinicians, especially those with no sound background in bioinformatics and who lack sufficient data analysis skills to interpret something from the plethora of cancer-specific data generated nowadays.

Key words: omics data; cancer biomarker; web server; computational resource; diagnosis; prognosis

Introduction

Genomic instability is often associated with the development of human diseases. These genomic instability events can occur at each step of genomic organization. Thus, understanding human health and disease require the proper investigation of molecular intricacy at genome organization levels such as the genome, proteome, epigenome, transcriptome, post-transcriptome and metabolome. With the development and advancements of next-generation sequencing (NGS) technologies, oncology research becomes data-driven. The data analysis at each genome organization level reveals that cancer is a heterogeneous and complex disease [1]. The analysis of generated data has completely revolutionized the cancer genomics field. Thus, in today's era, cancer genome analysis and clinical information have become a frontier in the management of cancer patients [2]. Therefore, integrating data generated at each genome level is essential to understand the complex nature of cancer and get a holistic overview of genomic instability events, which otherwise is not possible by single omics data analysis. In the recent decade, several clinical and preclinical studies showed the importance of data integration to get a clear and concise picture of the disease under investigation [3, 4]. In one study, researchers showed the importance of integrating proteomic data along with genomic and clinical data for the prioritization of driver genes in

Harpreet Kaur, PhD in Bioinformatics, Bioinformatics Centre, CSIR-Institute of Microbial Technology, Sector 39-A, Chandigarh-160036, India.
Rajesh Kumar, PhD in Life Sciences, Bioinformatics Centre, CSIR-Institute of Microbial Technology, Sector 39-A, Chandigarh-160036, India.
Anjali Lathwal, PhD in Computational Biology, Department of Computational Biology, Indraprastha Institute of Information Technology, New Delhi-110020, India.
Gajendra P.S. Raghava, PhD in Bioinformatics, Professor and Head, Department of Computational Biology, Indraprastha Institute of Information Technology, New Delhi-110020, India.

Table 1. Different types of omics techniques used in quantifying the genomic architecture of the human genome and their functionalities

Type of omics	Technique	Description
Genomics	WGS, WES	Used for variant identification at genome or exome level
Epigenomics	ChiP-Seq	Identification for DNA binding site, transcription factor
	DNase-Seq	Identification of regulatory elements
	ChiRP-Seq	Identification of ncRNA, lncRNA and their associated proteins
	WG-bisulphite	Identification of methylation sites in human genome
Transcriptomics	RNA-Seq	Identification of transcripts such as mRNA, miRNA
Proteomics	LC–MS/MS based	Quantify protein abundance within biological condition
	RRPA/SILAC based	Quantify protein abundance within biological condition
Metabolomics	LC–MS based	Identify and quantify metabolites involves in specific pathways

colorectal cancer [5]. The most convenient and widely used omics technologies to study genomic architecture are based on NGS and mass spectroscopy [6]. In a more precise way, NGS-based technology can be further categorized into several subtypes. The whole-genome sequencing is an NGS-based technology, particularly used for the identification of sequence variants in exome sequences [7]. The Chip-Seq, DNAse-Seq, FAIRE-Seq, are other NGS-based techniques used for the quantification of DNA-protein interaction and the identification of regulatory elements with the human genome [8]. Other NGS-based techniques are ChiRP-Seq and WG bisulfite sequencing, commonly employed to identify noncoding RNA and the methylation sites within the human genome, unraveling the epigenomic portion of genomic architecture [9]. Another variant of the NGS technique is the RNA-Seq, which is employed to quantify different kinds of miRNA and gene expression [8]. Thus we can conclude that NGS-based technologies are used to identify genomic alteration and variants at both coding and noncoding levels. In contrast to the NGS, the mass spectroscopy-based techniques are used to identify and quantify proteins in different subjects of interest [10]. The techniques, which are based on reverse-phase protein array and stable isotope labeling by amino acids in cell culture, are also another class of mass spectroscopy-based techniques used to quantify the protein molecules in the human genome [11]. All the different types of technologies that uncover the human genome's different omics layers are provided in Table 1.

These integrated approaches have also resulted in the development of several tools, resources and methods. These developed platforms provide a framework for genomic data integration, analysis, download, interpretation and visualization. Several review articles in the literature cover multidata integration and highlighted the importance of the same. The present review specifically lists the major single/multiomics databases present in the literature, including the resources that integrate data from various databases and data analysis servers employing the multiomics data for prognostic and predictive biomarkers identification. The major aim of this review article is to provide the scientific community a holistic view of the available resources for multiomics data integration and analysis to improve cancer therapeutics. We will specifically focus on the applicability of the available resources to understand the complex nature of human cancer and in the identification of various predictive and prognostic biomarkers. The schematic representation of the overall methodology is in Figure 1.

Resources on cancer genomics

The multiomics techniques can generate the data from each genomic hierarchy level ranging from genome to proteome. The annotation of genes, proteins and regulatory elements from various omics layers could serve as the basis for identifying disease-related outcomes. The data originated from the same set of samples or across the human population could help gain insight into the biological context of the disease onset and progression. With the advent of sequencing technologies, thousands of human genomes are sequenced. This will lead to the generation of large voluminous data. The generated multiomics data are thus stored in several dedicated repositories. There is a list of publicly available cancer-specific data resources in the literature. These cancer-specific online repositories of multiomics could play a key role in broadening our understanding of diseases related pathways and mechanisms. These web repositories could also provide a way to measure the altered molecular pattern of different molecular processes within the same cancer type or in a pan-cancer way. The data from these repositories allow the researchers to reinvestigate the data to gain meaningful insight into the disease etiology. For simplicity, we have categorized the different cancer-specific repositories into two types—primary and secondary. The Primary repositories are the storehouse of data generated from sequencing platforms and manual curation. The primary repositories contain information on multiomics aspects of cancer genome includes large genome consortium such as The Cancer Genome Atlas (TCGA), Gene Expression Omnibus (GEO), International Cancer Genome Consortium (ICGC), Encyclopedia of the Non-coding Elements of Human Genome (ENCODE), Sequence Read Archive (SRA), Cancer Cell Line Encyclopedia, etc. The GEO catalogs the data on gene expression profiling of the patients, genome methylation, genome variation and protein profiling studies of several diseases including cancer [12]. The TCGA is another such repository that provides comprehensive information on genomic, epigenomic, transcriptomics, clinical and proteomics information on 33 major human cancer types. The ICGC is another major cancer-specific data repository that stores both open access and controlled dataset on human cancer types. This portal offers various tools for data analysis and integration like simple gene-oriented queries and integrates genomic and clinical data. The key features of this resource include the comprehensiveness, high resolution, and quality of the data obtained from matched nontumor tissue, generation of complementary catalogs of transcriptomic and epigenomic datasets from the same tumors [13]. The SRA was established by the International Nucleotide Sequence Database Collaboration with the primary goals of including the storage of raw sequencing data, the alignment information from multiple high-throughput sequencing platforms and making this sequencing data easily available to the scientific community [14]. Besides the genomic and sequencing information, several other primary resources are also there in the literature. These

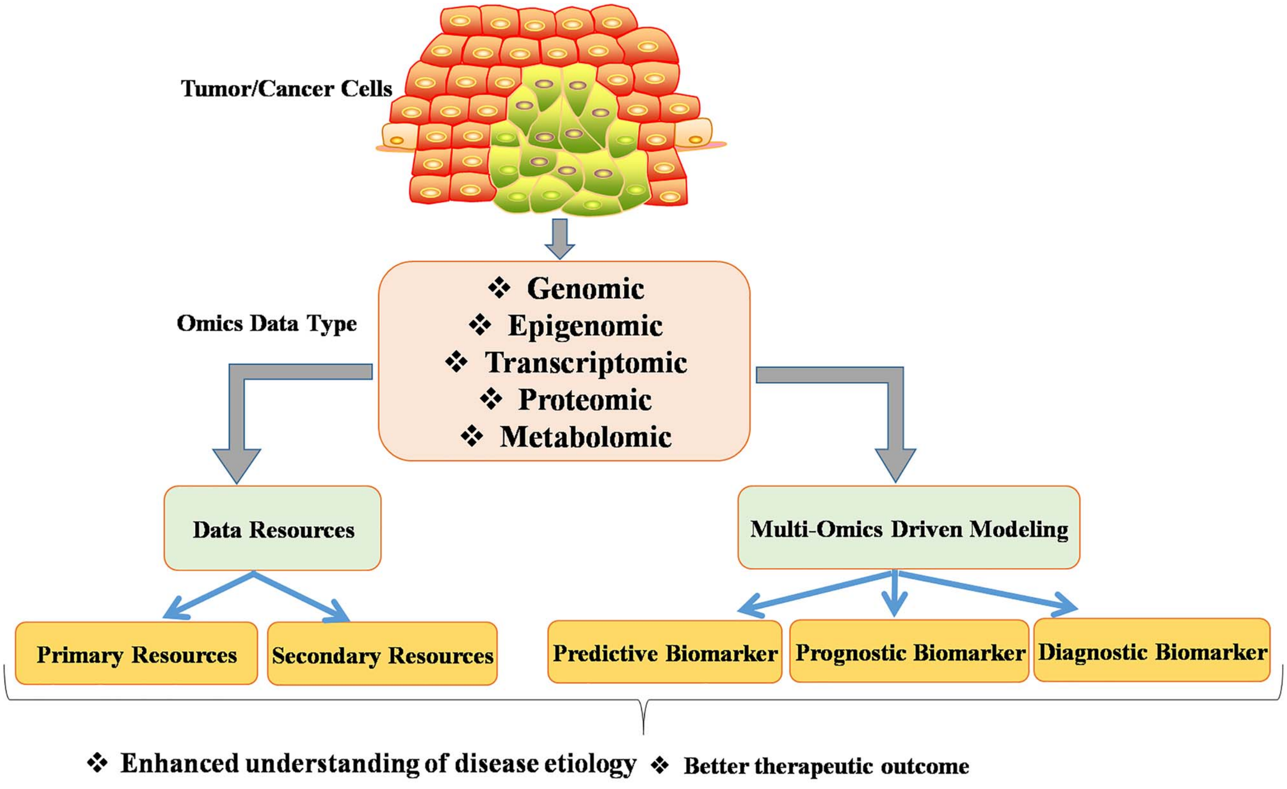

Figure 1. Schematic description of the overall methodology and workflow of the review article.

resources are manually curated by the team of experts and provide extensive information regarding cancer phenotype and targeted therapeutic approaches. The Catalog of Somatic Mutation in Cancer (COSMIC) [15] is one such significant effort that provides somatic mutation information on almost every human gene. COSMIC is developed by many experts and contains manually curated information from 27 724 research publications, TCGA and ICGC [15]. The latest release of COSMIC (v92, release 27 august 2020) catalogs 37 288 077 coding mutations, 9 215 470 coding region variants, 15 642 672 noncoding region variants, 19 369 fusion transcripts, 1 207 190 copy number variants, 7 930 489 methylated CpG. In addition to the COSMIC, several other groups worldwide also develop some manually curated repositories that provide comprehensive information on cancer progression genomics. Notable examples include CancerPDF—a repository of cancer-related endogenous peptides that are detected in the human biofluids; colorectal cancer biomarker database—catalogs 870 biomarker information from 1115 research articles; Liverome—holds 143 liver cancer-specific signatures containing 6927 genes extracted from 98 research publications. Further, in the recent past, circular RNAs emerged as noninvasive biomarker candidates among variety of diseases including cancer. Hence, various web resources, such as CircNet, CircRiC ExoRBase, Circ2Disease, Circ2Traits, LncRNADisease 2.0, MiOncoCirc, etc. were developed to maintain the extensive information for CircRNAs and their association with multiomics layers and diseases [16–22]. These CircRNA resources and tools for the identification these CircRNA discussed in more details by Zheng *et al.* elsewhere [23, 24]. In addition to the primary resources such as GDC Data portal, ICGC ArrayExpress, UCSC Genome Browser, T3CA, LinkedOmics, COSMIC, GDSC, CCLE, etc. [25–40], there exist several secondary resources like Liverome, CancerPDF, CancerPPD, CBD, cBioPortal, HCMDB, CancerDR, SomamiR, OncoMX, ResMarkerDB, CancerLivER, HCCdb, DBMHCC, ApoCanD, etc. [41–56], in the literature. Secondary resources are developed by extracting and integrating data from the primary resources. These include—human cancer metastasis database, CancerEnD, CancerDR, CancerMIRNome, OncoMX, etc. The complete list of primary and secondary cancer-related resources with their brief description, URLs, Pro/Cons with references is in Supplementary Table S1.

Biomarkers tools for cancer research

Multiomics platforms are generating an enormous amount of data. The integration of data to mine the biologically meaningful insight is the major challenge faced by the researchers and clinicians. In this regard, a growing number of researchers worldwide continuously integrated the various omics dataset to get a better insight into the disease etiology [57–66]. This data integration approach can uncover the hidden dynamic properties of cancer cells that otherwise would be impossible with the static or single omics data analysis. In the past, several tools are developed by the researchers by employing the multiomics data integration approach, which finds use in indentifying the cancer-driving genes [67], predicting the survival of patients, predicting the success rate of cancer immunotherapy [68, 69]. The researchers have extensively utilized the omics profile of cancer patients to identify potential biomarkers for diagnosis and prognosis purposes [57, 70]. Many of these studies also developed webtools based on identified omics biomarkers to predict the status of the tumor. Several other powerful bioinformatics tools like KMPlotter, Gene Expression Profiling Interactive Analysis, Oncomine, etc. are also developed by the researchers. These resources were developed to investigate the publicly available omics datasets for advancement in oncology research. However, some of these tools are developed on single omics datasets and often require a tedious registration process. Thus overcoming such challenges is necessary for improving the cancer therapeutics strategy. In this regard, Yan *et al.* developed a web-resource for lung cancer patients' survival prediction by integrating 5245 samples from TCGA, GEO and other publicly available sources [71]. Dong *et al.* also developed a tool, OSgbm [72], by integrating datasets from seven resources, namely TCGA, GEO and Chinese cancer databases. The researchers have also integrated data in one of the studies to unravel the hidden heterogeneity of the cancer types [60], thus providing subtype-specific biomarkers. The data from all these reports highlight that integrating multiomics data can help improve therapeutics for cancer management.

In addition to single gene-based biomarkers, researchers also developed multigene-based biomarkers tools to capture more tumor heterogeneity. In one such study, Kaur *et al.* have identified a set of three genes based biomarkers (*CLEC1B, PRC1* and *FCN3*) that have high diagnostic potential with more than 90% accuracy employing various statistical and machine learning algorithms. The developed machine learning model has been implemented in a web-tool, i.e. HCCPred [63]. CancerCSP is another web-tool that can classify the patients into the early and late stages of renal cell carcinoma based on the expression profile of 64 and 38 genes [73]. Kumar *et al.* in one study, developed the tool by integrating a dataset from TCGA and other literature studies to find the prognostic potential of enhancer elements for 18 cancer types [4]. These online free-to-use tools and web-resources help clinicians and researchers discover several prognostic and diagnostic markers that are ultimately beneficial for cancer research. The tools developed further can be classified into Diagnostic and Prognostic biomarker tools [e.g. OSluca, OSgbm, CancerCSP, CancerLSP, CancerUBM, BBCancer, OScc, OSbrca, PROGgene, SurvExpress, PrognoScan, GSCAlite, CaPSSA, MEXPRESS, PROGmiR, SurvMicro, OncoLnC, TCPAv3.0, TRGAted [71–90]) or Predictive (e.g. CancerDP, CancerTOPE, SCLC-CellMiner [91–93]) biomarker tools. The tools that can help identify these biomarkers are of great significance in guiding the clinical treatment, elucidation of the mechanism of tumorigenesis, and offers an opportunity to clinicians for targeted therapy. Supplementary Table S2 provides a brief description of diagnostic, prognostic and predictive tools used in cancer research.

Miscellaneous tools for cancer research

The publically available resources such as TCGA, GEO provides a huge amount of data on genomics, proteomics, etc. Thus, these data provide unparalleled opportunities for researchers and clinicians to explore gene function analysis, biological mechanism discussion and target identification. Analysis of the generated omics dataset provides several biomarkers in clinical use and unravels the cancer genome's hidden targets. Analyzing and integrating such complex datasets is also a tedious task and demands a bioinformatics expertise person. Researchers without strong computing potential and bioinformatics background often find difficulty in analyzing and interpreting the data. Thus, if genomic studies' full potential has to be utilized in clinics, there should some alternative tools for easy data visualization, interpretation and analysis. Using such tools, researchers who do not belong to the bioinformatics class can also ask specific questions and generate a testable hypothesis. Thus to aid the scientific community, several comprehensive analysis tools, i.e., GEPIA2, SCUDO, ChIP-Array2, DeAnnCNV, Sniplay3, varFish, Oviz-Bio, Cancer3D 2.0, FireBrowse [94–102] are also developed by the researchers. The complex genomic analysis can be done by simple clicks using these servers and tools. Supplementary Table S3 catalogs some of the available resources and tools used for easy data visualization and analysis.

Comparative analysis of tools and algorithms for cancer research

The rapid growth in the multiomics data opens a platform for researchers in aggregating, integrating and analysis of the data to drive novel targets and advances in clinical research. Thus to improve biomedical knowledge, teams of research scientists with diverse backgrounds have worked hard to develop advanced methods and tools for efficient data management, handling and interpretation. Given the wide spectrum of developed tools (refers to Supplementary Table S1–S3) with the varying approach in data integration, feature selection, clustering, data interpretation, and analysis, a detailed comparison of developed tools in the context of the same data set could be very useful for benchmarking the performance and evaluation of their suitability in the clinics. Several studies are available in the literature which performed the comparison of the clustering and feature selection methods used in these tools for multiomics data integration [103, 104]. The researchers in one study suggest that SNF is the most robust feature selection step among the other available such as MCCA, MFA, MCIA and JIVE [104]. Their observation is based on the comparative analysis of all the methods based on three real datasets with a varying number of parameters such as feature selection, noise ratio, signal strength and training parameters. Another study performed a comparative analysis of available methods for clustering analysis. This study compared the clustering performance of six available methods on the simulated datasets and conclude that BCC (Bayesian approach) showed the highest accuracy [104]. The BCC methods can correctly identify data specific structures across the datasets. The study by Rappoport *et al.* also compared the six clustering algorithms such as LRAcluster, K-means, SNF, multiNMS, PINSPlus, iclusterBayes, spectral clustering, rMKL-LPP and MCCA on the cancer multiomics dataset from the TCGA. They conclude that for the clustering of multiomics gene expression, miRNA expression, and DNA methylation the rMKL-LPP, MCCA and multiNMF outperformed other available algorithms in terms of clinical subtype-specific enrichment [105]. Several gene expression-based computational tools have been developed for the prediction of survival outcomes in the patients. Each tool utilizes different inclusion and exclusion criteria with little or no overlap between the patient cohorts. This will leads to irreproducibility among the results of the published studies and thus limit their use in clinical settings. Altman *et al.* in one study developed an intuitive algorithm PCM-SABRE that compares and benchmarks the gene expression-based survival prediction using various machine learning algorithms [106]. This study showcased the power of different feature selection and machine learning algorithms to improve the existing expression-based prediction models. PNN and LR machine learning algorithms perform better than other algorithms and the ANOVA feature

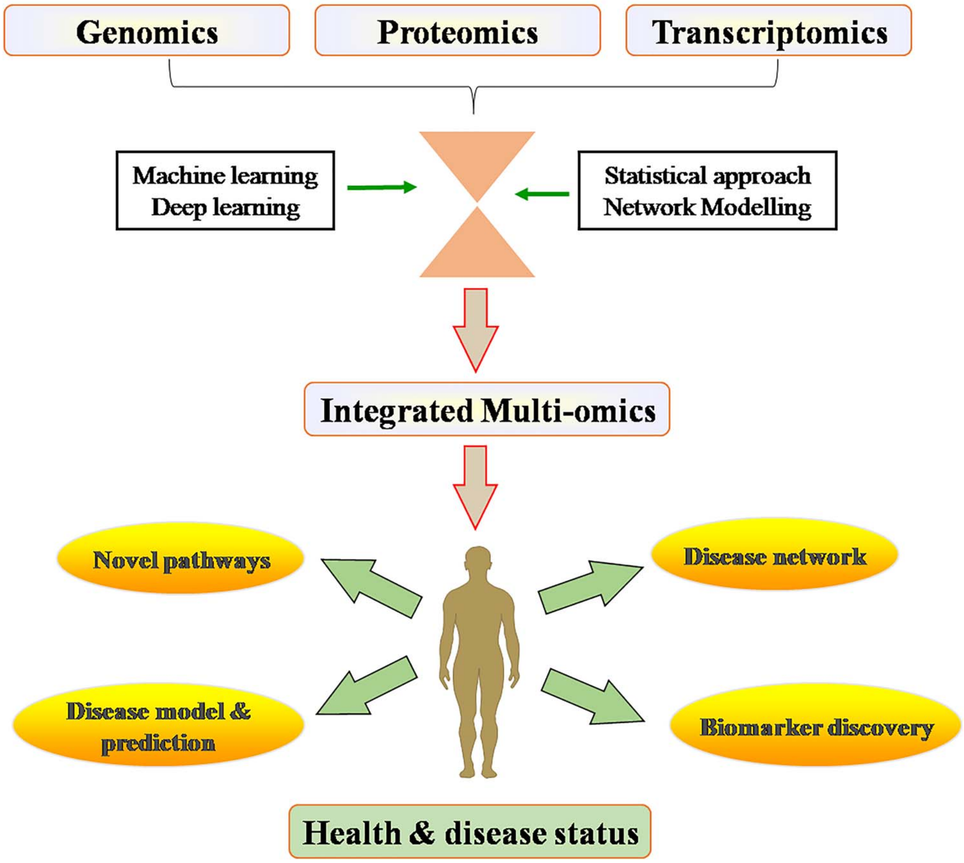

Figure 2. Depiction of workflow and application of multiomics data integration in cancer research.

selection method achieved higher accuracy as compared to other feature selection methods for breast cancer survival prediction. Overall, a large number of studies compared and benchmarked different omics-related tools and algorithms. However, it is not wrong to say that the data and features employed in any model are the major deciding and driving forces leading to the better performance of one algorithm and tool over another.

Application of tools in current cancer research

The computational tools which integrate, analyze and derive useful insights from the available genomic or public datasets are continuously being released in the literature. Herein, we discuss some potential applications of the available computational tools with their technical details. The tools discussed are organized based on their applicability in addressing the biological question of interest. For the sake of simplicity, we have categorized the biological question of interest in two broad categories—(I) identification of disease subtype-specific features, (II) identification of disease-specific biomarkers for diagnostic, prognostic and therapeutic purposes. A brief description of the applicability of the developed computation tools is provided in Figure 2.

Identification of disease subtype-specific features

Cancer is a heterogeneous disease in terms of disease etiology and progression [106]. It shows a remarkable degree of heterogeneity at various levels like intratumor, intertumor and also at the patient level. Several other factors such as environment, lifestyle also contributes to the varying degree of disease heterogeneity. Thus, it is of utmost importance to identify the disease subtype-specific biomarkers that may be beneficial in suggesting and designing interventions for patients in a more personalized and subtype-specific manner [106]. PINSPlus is one such available method that can classify the patients into different subtypes using multiomics data [107]. PINSPlus uses the similarity index-based algorithm to cluster the patients into different disease subtypes. Based on the hierarchical structure search, it can also be used for the identification of subgroups within the subtypes. For example, PINPlus was applied on 34 omic data and two METABRIC data for breast cancer study for the identification of subtype-specific differences in survival of the patients. The tool was able to identify the subtypes for 27 datasets out of 36 with a significant *P*-value for the difference in survival among the subtypes. Several similar studies depict the importance of different multiomics tools in the advancement of disease subtyping.

Identification of disease-specific biomarkers for diagnostic, prognostic and therapeutic purposes

Biomarkers are often considered as a gold standard molecular footprint for revealing the condition of living cells. They can provide accurate information on the connected pathway and flow of information in a cell and thus have the potential to reveal disease etiology. The multiomics data integration approach unveils an innovative platform for the identification of disease etiology specific biomarkers, which can further be used in clinics for diagnostic, prognostic and therapeutic purposes. NetICS is one such method available for the integration of multiomics data and prioritization of cancer disease-specific genes [108]. This tool is developed by integrating mutation, miRNA, mRNA and CNV data on five different cancer types available in TCGA. The developed method was able to identify the frequent and infrequent altered genes and thus can also be used in the ranking of genes in terms of their diagnostic potential. CancerSPP is another freely available computational tool for the prediction of the progression of cancer by conducting the integrative analysis of multiomics data namely mRNA, miRNA and methylation status of skin cutaneous melanoma patients from the TCGA. Various machine learning techniques have been employed for the development of a computational pipeline that can classify samples into primary and metastatic categories independently. In addition to gene-based markers, researchers also developed computational tools for the risk stratification of colorectal patients based on their protein expression profiles [61]. The weight factor for each protein was taken into consideration for the development of the prediction model to classify the patients into the various risk groups. These tools can help in reducing the overall burden of cancer deaths worldwide by acting as a platform to aid in the timely diagnosis, better prognosis and suggesting advanced personalized disease-specific therapeutic options.

Future direction

Cancer exploits different mechanisms to alter the normal physiology and cellular processes to their benefit to activate various immune escape pathways leading to cancer onset, progression and therapeutic failures. The decades of global collaborative clinical and preclinical research significantly advanced our knowledge regarding disease diagnosis, treatment and management which resulted in improved outcomes in the patients. This significant increase in patient care is driven by rapid advancement in the field of genomics, bioinformatics, sequencing and imaging technologies with properly established electronic health records. The rapid increase in genomics and related technological fields lead to the generation of an enormous amount of data that is freely available in numerous repositories and a large number of analytical tools exploit this data to solve the particular problem of interest. The stored data in specialized bioinformatics resources such as TCGA, ICGC, COSMIC, ENCODE, MethyCancer have been continuously utilized by researchers and clinicians for biomarker discovery. Resources such as CanSAR, GDSC supports clinical cancer research in the drug discovery process. Despite the importance of developed tools, the shift from bench to bedside remains a challenging task. One of the major challenges that are still needed to be resolved is the heterogeneity of the tumor cells leading to varied responses to anticarcinogen/therapy with a similar response. This obstacle can be overcome by initiating the personnel genomics projects in large human populations of different geographical locations. This will helps in the decoding of the personalized mutational fingerprints and molecular makeup of the cancer types. The developed tools on such large-scale datasets can better mimic the local and systemic tumor microenvironment and thus can be utilized in better management of therapies. The prolonged survival of cancer patients with checkpoint inhibitors and immunotherapies are still restricted to only a minuscule set of patients. The low survival among the patient is because there is no definite biomarker that can recognize the patient subset and can subsequently optimize the delivery and selection process. To achieve long-term survival, a combination of drugs targeting several molecular perturbations and cancer driver mutations might be needed. There is a need for the development of computational models that can help in predicting the drug response by analyzing the patient genomic and mutational profile in a more disease- and subtype-specific manner. Also, advanced computational and statistical tools should be developed which aims at better data management, integration and analysis to establish a strong causal relationship between clinical data, genomic discovery and overall patient care.

Conclusion

The advancement in omics technologies results in the generation of enormous data in the field of biomedical research. The availability of such a vast amount of data provides an opportunity to investigate and derive significance from this data for complex pathological conditions like cancer. Several resources manage the multiomics data of cancer. One of the advantages of multiomics data is that it can provide a holistic and broader picture of cancer compared to the single omics layer. It can help us identify better biomarkers for diagnosis, prognosis and predicting the treatment with high precision. In this review, first, we have provided an overview regarding multiomics data types, including genomics, proteomics, transcriptomics, epigenomics and metabolomics. Then, we provided a brief introduction to the major web resources that manage multiomics data for cancer. We divided resources into primary and secondary resources based on the source of the data. Next, we provided an overview of the methods and tools developed to explore and integrate multiomics data. We listed various tools based on machine learning, deep learning, survival analysis algorithms and different packages employing multiomics data to identify biomarkers for specific malignant conditions. Here, we divided these tools based on applying the method, i.e. diagnostic, prognostic, predictive and precision medicine. We have tried to cover major resources that can provide insight into the application of multiomics data in the biomarker discovery for cancer. The generated multiomics data have dramatically improved our understanding of cancer. However, the generated dataset adds another layer of difficulty for researchers in integrating and interpreting the data. The developed tools can assist clinicians and researchers in developing biomarkers, predicting the response to therapy, assessing risk scores, etc. This way, the developed tools help clinicians design and guide therapy to the patients. However, the developed tools also suffer from several limitations: integrating cancer tissue images, multinetwork model constructions, genomic pathway and metabolite information, etc. Thus, they need to be improved soon. After overcoming the above limitations, we hope that the bioinformatics tools may open a new avenue for biomarker discovery and better patient/healthcare management.

Key points

- With the advent of technology, there is an enormous generation of multiomics data of patients including genomics, proteomics, transcriptomics, epigenomics and metabolomics.
- A number of web resources developed to manage this vast amount of multiomics data for cancer samples.

We split these resources into primary and secondary resources based on the source of the data.

- A large number of studies identified potential omics biomarker candidates for specific malignant condition by exploring and integrating multiomics data of patients employing various machine learning, deep learning techniques, and survival analysis algorithms, etc.
- Various web-tools were developed to predict the tumor status of the samples based on identified omics biomarkers implementing different bioinformatics approaches. We categorized these tools mainly into four categories based on the application, i.e. diagnostic, prognostic, predictive and precision medicine.
- We have tried to cover major resources that can provide insight into the application of multiomics data in the biomarker discovery for cancer. The generated multiomics data has dramatically improved our understanding of cancer.

Supplementary data

Supplementary data are available online at http://bib.oxfordjournals.org/.

Author's contribution

Harpreet Kaur, Rajesh Kumar, Anjali Lathwal and GPS Raghava performed conception and design of the study; Harpreet Kaur, Anjali Lathwal, Rajesh Kumar and GPS Raghava did analysis and interpretation of data; Rajesh Kumar and Anjali Lathwal did the preparation of figures and tables; Harpreet Kaur, Rajesh Kumar, Anjali Lathwal and GPS Raghava performed the writing, reviewing and draft preparation of the manuscript. All authors have read and approved the manuscript.

Funding Information

There is no funding available for this article.

Acknowledgement

The authors acknowledge funding from J. C. Bose National Fellowship from Department of Science & Technology (DST), India with Grant Number SRP076. H.K. and R.K. are thankful to Council of Scientific and Industrial Research (CSIR), India, and A.L. is thankful to UGC, India for providing fellowship. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Conflicts of Interest

The authors declare no financial and non-financial conflict of interest.

References

1. Blackadar CB. Historical review of the causes of cancer. *World J Clin Oncol* 2016;**7**:54–86.
2. Lathwal A, Kumar R, Raghava GPS. Computer-aided designing of oncolytic viruses for overcoming translational challenges of cancer immunotherapy. *Drug Discov Today* 2020;**25**:1198–205.
3. Subramanian I, Verma S, Kumar S, *et al*. Multi-omics data integration, interpretation, and its application. *Bioinform Biol Insights* 2020;**14**:1177932219899051.
4. Kumar R, Lathwal A, Kumar V, *et al*. CancerEnD: a database of cancer associated enhancers. *Genomics* 2020;**112**:3696–702.
5. Zhang B, Wang J, Wang X, *et al*. Proteogenomic characterization of human colon and rectal cancer. *Nature* 2014;**513**:382–7.
6. Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet* 2016;**17**:333–51.
7. Petersen BS, Fredrich B, Hoeppner MP, *et al*. Opportunities and challenges of whole-genome and -exome sequencing. *BMC Genet* 2017;**18**:108–120.
8. Das T, Andrieux G, Ahmed M, *et al*. Integration of online omics-data resources for cancer research. *Front Genet* 2020;**11**:578345.
9. Nagarajan RP, Fouse SD, Bell RJA, *et al*. Methods for cancer epigenome analysis. *Adv Exp Med Biol* 2013;**754**: 313–38.
10. Prieto DA, Johann DJ, Wei BR, *et al*. Mass spectrometry in cancer biomarker research: a case for immunodepletion of abundant blood-derived proteins from clinical tissue specimens. *Biomark Med* 2014;**8**:269–86.
11. Bohnenberger H, Ströbel P, Mohr S, *et al*. Quantitative mass spectrometric profiling of cancer-cell proteomes derived from liquid and solid tumors. *J Vis Exp* 2015;**96**: e52435.
12. Barrett T, Wilhite SE, Ledoux P, *et al*. NCBI GEO: archive for functional genomics data sets - update. *Nucleic Acids Res* 2013;**41**:D991–5.
13. Hudson TJ, Anderson W, Aretz A, *et al*. International network of cancer genome projects. *Nature* 2010;**464**: 993–8.
14. Leinonen R, Sugawara H, Shumway M. The sequence read archive. *Nucleic Acids Res* 2011;**39**:D19–21.
15. Bamford S, Dawson E, Forbes S, *et al*. The COSMIC (catalogue of somatic mutations in cancer) database and website. *Br J Cancer* 2004;**91**:355–8.
16. Yao D, Zhang L, Zheng M, *et al*. Circ2Disease: a manually curated database of experimentally validated circRNAs in human disease. *Sci Rep* 2018;**8**:11018.
17. Ghosal S, Das S, Sen R, *et al*. Circ2Traits: a comprehensive database for circular RNA potentially associated with disease and traits. *Front Genet* 2013;**4**:283.
18. Vo JN, Cieslik M, Zhang Y, *et al*. The landscape of circular RNA in cancer. *Cell* 2019;**176**:869–881.e13.
19. Li S, Li Y, Chen B, *et al*. ExoRBase: a database of circRNA, lncRNA and mRNA in human blood exosomes. *Nucleic Acids Res* 2018;**46**:D106–12.
20. Bao Z, Yang Z, Huang Z, *et al*. LncRNADisease 2.0: an updated database of long non-coding RNA-associated diseases. *Nucleic Acids Res* 2019;**47**:D1034–7.
21. Liu YC, Li JR, Sun CH, *et al*. CircNet: a database of circular RNAs derived from transcriptome sequencing data. *Nucleic Acids Res* 2016;**44**:D209–15.
22. Ruan H, Xiang Y, Ko J, *et al*. Comprehensive characterization of circular RNAs in ∼ 1000 human cancer cell lines. *Genome Med* 2019;**11**:55.

23. Zeng X, Lin W, Guo M, *et al*. A comprehensive overview and evaluation of circular RNA detection tools. *PLoS Comput Biol* 2017;**13**:e1005420.
24. Vromman M, Vandesompele J, Volders P-J. Closing the circle: current state and perspectives of circular RNA databases. *Brief Bioinform* 2020;**22**:288–297.
25. Jensen MA, Ferretti V, Grossman RL, *et al*. The NCI genomic data commons as an engine for precision medicine. *Blood* 2017;**130**:453–9.
26. Zhang J, Baran J, Cros A, *et al*. International cancer genome consortium data portal-a one-stop shop for cancer genomics data. *Database* 2011;**2011**:bar026.
27. Brazma A, Parkinson H, Sarkans U, *et al*. ArrayExpress - a public repository for microarray gene expression data at the EBI. *Nucleic Acids Res* 2003;**31**:68–71.
28. Zhu J, Sanborn JZ, Benz S, *et al*. The UCSC cancer genomics browser. *Nat Methods* 2009;**6**:239–40.
29. Feng X, Li L, Wagner EJ, *et al*. TC3A: the cancer 3′ UTR atlas. *Nucleic Acids Res* 2018;**46**:D1027–30.
30. Lee M, Lee K, Yu N, *et al*. ChimerDB 3.0: an enhanced database for fusion genes from cancer transcriptome and literature data mining. *Nucleic Acids Res* 2017;**45**: D784–9.
31. Vasaikar SV, Straub P, Wang J, *et al*. LinkedOmics: Analyzing multi-omics data within and across 32 cancer types. *Nucleic Acids Res* 2018;**46**:D956–63.
32. Ryan M, Wong WC, Brown R, *et al*. TCGASpliceSeq a compendium of alternative mRNA splicing in cancer. *Nucleic Acids Res* 2016;**44**:D1018–22.
33. Halling-Brown MD, Bulusu KC, Patel M, *et al*. canSAR: an integrated cancer public translational research and drug discovery resource. *Nucleic Acids Res* 2012;**40**:D947–56.
34. Packer BR, Yeager M, Burdett L, *et al*. SNP500Cancer: a public resource for sequence validation, assay development, and frequency analysis for genetic variation in candidate genes. *Nucleic Acids Res* 2006;**34**:D617–21.
35. Samur MK, Yan Z, Wang X, *et al*. canEvolve: a web portal for integrative Oncogenomics. *PLoS One* 2013;**8**:e56228.
36. Liu SH, Shen PC, Chen CY, *et al*. DriverDBv3: a multi-omics database for cancer driver gene research. *Nucleic Acids Res* 2020;**48**:D863–70.
37. Ru B, Sun J, Tong Y, *et al*. CR2Cancer: a database for chromatin regulators in human cancer. *Nucleic Acids Res* 2018;**46**:D918–24.
38. Yang W, Soares J, Greninger P, *et al*. Genomics of drug sensitivity in cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Res* 2013;**41**:D995–61.
39. Barretina J, Caponigro G, Stransky N, *et al*. The cancer cell line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* 2012;**483**:603–7.
40. Sun X, Pittard WS, Xu T, *et al*. Omicseq: a web-based search engine for exploring omics datasets. *Nucleic Acids Res* 2017;**45**:W445–52.
41. Lee L, Wang K, Li G, *et al*. Liverome: A curated database of liver cancer-related gene signatures with self-contained context information. *BMC Genomics*. 2011; **12**: Suppl 3(Suppl 3):S3.
42. Bhalla S, Verma R, Kaur H, *et al*. CancerPDF: a repository of cancer-associated peptidome found in human biofluids. *Sci Rep* 2017;**7**:1511.
43. Tyagi A, Tuknait A, Anand P, *et al*. CancerPPD: a database of anticancer peptides and proteins. *Nucleic Acids Res* 2015;**43**:D837–43.
44. Zhang X, Sun XF, Cao Y, *et al*. CBD: a biomarker database for colorectal cancer. *Database* 2018;**2018**:bay046.
45. Cerami E, Gao J, Dogrusoz U, *et al*. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov* 2012;**2**:401–4.
46. Zheng G, Ma Y, Zou Y, *et al*. HCMDB: the human cancer metastasis database. *Nucleic Acids Res* 2018;**46**:D950–5.
47. Kumar R, Chaudhary K, Gupta S, *et al*. CancerDR: cancer drug resistance database. *Sci Rep* 2013;**3**:1445.
48. Bhattacharya A, Cui Y. SomamiR 2.0: a database of cancer somatic mutations altering microRNA-ceRNA interactions. *Nucleic Acids Res* 2016;**44**:D1005–10.
49. Dingerdissen HM, Bastian F, Vijay-Shanker K, *et al*. OncoMX: a knowledgebase for exploring cancer biomarkers in the context of related cancer and healthy data. *JCO Clin Cancer Informatics* 2020;**4**:210–20.
50. Pérez-Granado J, Piñero J, Furlong LI. ResMarkerDB: a database of biomarkers of response to antibody therapy in breast and colorectal cancer. *Database* 2019;**2019**: baz060.
51. Kaur H, Bhalla S, Kaur D, *et al*. CancerLivER: a database of liver cancer gene expression resources and biomarkers. *Database (Oxford)* 2020;**2020**:baaa012.
52. Lian Q, Wang S, Zhang G, *et al*. HCCDB: a database of hepatocellular carcinoma expression atlas. Genomics. *Proteomics Bioinforma* 2018;**16**:269–75.
53. Chu YW, Chien CH, Sung MI, *et al*. DBMHCC: a comprehensive hepatocellular carcinoma (HCC) biomarker database provides a reliable prediction system for novel HCC phosphorylated biomarkers. *PLoS One* 2020;**15**:e0234084.
54. Kumar R, Raghava GPS. ApoCanD: database of human apoptotic proteins in the context of cancer. *Sci Rep* 2016;**6**:20797.
55. Nagpal G, Sharma M, Kumar S, *et al*. PCMdb: pancreatic cancer methylation database. *Sci Rep* 2014;**4**:4197.
56. Agarwal SM, Raghav D, Singh H, *et al*. CCDB: a curated database of genes involved in cervix cancer. *Nucleic Acids Res* 2011;**39**:D975–9.
57. Mason MJ, Schinke C, Eng CLP, *et al*. Multiple myeloma DREAM challenge reveals epigenetic regulator PHF19 as marker of aggressive disease. *Leukemia* 2020;**34**:1866–74.
58. Vidyarthi A, Agnihotri T, Khan N, *et al*. Predominance of M2 macrophages in gliomas leads to the suppression of local and systemic immunity. *Cancer Immunol Immunother* 2019;**68**:1995–2004.
59. Bhalla S, Kaur H, Kaur R, *et al*. Expression based biomarkers and models to classify early and late-stage samples of papillary thyroid carcinoma. *PLoS One* 2020;**15**:e0231629.
60. Lathwal A, Kumar R, Arora C, *et al*. Identification of prognostic biomarkers for major subtypes of non-small-cell lung cancer using genomic and clinical data. *J Cancer Res Clin Oncol* 2020;**146**:2743–52.
61. Lathwal A, Arora C, Raghava GPS. Prediction of risk scores for colorectal cancer patients from the concentration of proteins involved in mitochondrial apoptotic pathway. *PLoS One* 2019;**14**:e0217527.
62. Arora C, Kaur D, Lathwal A, *et al*. Risk prediction in cutaneous melanoma patients from their clinico-pathological features: superiority of clinical data over gene expression data. *Heliyon* 2020;**6**:e04811.
63. Kaur H, Dhall A, Kumar R, *et al*. Identification of platform-independent diagnostic biomarker panel for hepatocellular carcinoma using large-scale transcriptomics data. *Front Genet* 2019;**10**:1306.

64. Dhall A, Patiyal S, Kaur H, *et al*. Computing skin cutaneous melanoma outcome from the HLA-alleles and clinical characteristics. *Front Genet* 2020;**11**:221.
65. Bhalla S, Kaur H, Dhall A, *et al*. Prediction and analysis of skin cancer progression using genomics profiles of patients. *Sci Rep* 2019;**9**:15790.
66. Kaur H, Bhalla S, Garg D, *et al*. Analysis and prediction of cholangiocarcinoma from transcriptomic profile of patients. *J Hepatol* 2020;**73**:S16–7.
67. Hua X, Xu H, Yang Y, *et al*. DrGaP: a powerful tool for identifying driver genes and pathways in cancer sequencing studies. *Am J Hum Genet* 2013;**93**:439–51.
68. Palmisano A, Krushkal J, Li MC, *et al*. Bioinformatics tools and resources for cancer immunotherapy study. *Methods Mol Biol* 2020;**2055**:649–78.
69. Lathwal A, Kumar R, Raghava GPS. OvirusTdb: a database of oncolytic viruses for the advancement of therapeutics in cancer. *Virology* 2020;**548**:109–16.
70. Chaudhary K, Poirion OB, Lu L, *et al*. Deep learning–based multi-omics integration robustly predicts survival in liver cancer. *Clin Cancer Res* 2018;**24**:1248–59.
71. Yan Z, Wang Q, Lu Z, *et al*. OSluca: an interactive web server to evaluate prognostic biomarkers for lung cancer. *Front Genet* 2020;**11**:420.
72. Dong H, Wang Q, Li N, *et al*. OSgbm: an online consensus survival analysis web server for glioblastoma. *Front Genet* 2020;**10**:1378.
73. Bhalla S, Chaudhary K, Kumar R, *et al*. Gene expression-based biomarkers for discriminating early and late stage of clear cell renal cancer. *Sci Rep* 2017;**7**:44997.
74. Kaur H, Bhalla S, Raghava GPS. Classification of early and late stage liver hepatocellular carcinoma patients from their genomics and epigenomics profiles. *PLoS One* 2019;**14**:e0221476.
75. Bhalla S, Chaudhary K, Gautam A, *et al*. A web bench for analysis and prediction of oncological status from proteomics data of urine samples. *bioRxiv* 2018;315564May 15, 2018. doi: 10.1101/315564 preprint: not peer reviewed.
76. Zuo Z, Hu H, Xu Q, *et al*. BBCancer: an expression atlas of blood-based biomarkers in the early diagnosis of cancers. *Nucleic Acids Res* 2020;**48**:D789–96.
77. Wang Q, Zhang L, Yan Z, *et al*. OScc: an online survival analysis web server to evaluate the prognostic value of biomarkers in cervical cancer. *Future Oncol* 2019;**15**:3693–9.
78. Yan Z, Wang Q, Sun X, *et al*. OSbrca: a web server for breast cancer prognostic biomarker investigation with massive data from tens of cohorts. *Front Oncol* 2019;**9**:1349.
79. Goswami CP, Nakshatri H. PROGgene: gene expression based survival analysis web application for multiple cancers. *J Clin Bioinforma* 2013;**3**:22.
80. Aguirre-Gamboa R, Gomez-Rueda H, Martínez-Ledesma E, *et al*. SurvExpress: an online biomarker validation tool and database for cancer gene expression data using survival analysis. *PLoS One* 2013;**8**:e74250.
81. Gentles AJ, Newman AM, Liu CL, *et al*. The prognostic landscape of genes and infiltrating immune cells across human cancers. *Nat Med* 2015;**21**:938–45.
82. Mizuno H, Kitada K, Nakai K, *et al*. PrognoScan: a new database for meta-analysis of the prognostic value of genes. *BMC Med Genomics* 2009;**2**:18.
83. Liu CJ, Hu FF, Xia MX, *et al*. GSCALite: a web server for gene set cancer analysis. *Bioinformatics* 2018;**34**:3771–2.
84. Jang Y, Seo J, Jang I, *et al*. CaPSSA: visual evaluation of cancer biomarker genes for patient stratification and survival analysis using mutation and expression data. *Bioinformatics* 2019;**35**:5341–3.
85. Koch A, Jeschke J, Van Criekinge W, *et al*. MEXPRESS update 2019. *Nucleic Acids Res* 2019;**47**:W561–5.
86. Goswami CP, Nakshatri H. PROGmiR: a tool for identifying prognostic miRNA biomarkers in multiple cancers using publicly available data. *J Clin Bioinforma* 2012;**2**:23.
87. Aguirre-Gamboa R, Trevino V. SurvMicro: assessment of miRNA-based prognostic signatures for cancer clinical outcomes by multivariate survival analysis. *Bioinformatics* 2014;**30**:1630–2.
88. Anaya J. OncoLnc: linking TCGA survival data to mRNAs, miRNAs, and lncRNAs. *Peer J Computer Science*. 2016;**2**: e67.
89. Chen MJM, Li J, Wang Y, *et al*. TCPA v3.0: an integrative platform to explore the pan-cancer analysis of functional proteomic data. *Mol Cell Proteomics* 2019;**18**: S15–25.
90. Zhang W, Borcherding N, Bormann NL, *et al*. TRGAted: a web tool for survival analysis using protein data in the cancer genome atlas. *F1000Research* 2018;**7**:1235.
91. Gupta S, Chaudhary K, Kumar R, *et al*. Prioritization of anticancer drugs against a cancer using genomic features of cancer cells: a step towards personalized medicine. *Sci Rep* 2016;**6**:23857.
92. Gupta S, Chaudhary K, Dhanda SK, *et al*. A platform for designing genome-based personalized immunotherapy or vaccine against cancer. *PLoS One* 2016;**11**:e0166372.
93. Tlemsani C, Pongor L, Elloumi F, *et al*. SCLC-CellMiner: a resource for small cell lung cancer cell line genomics and pharmacology based on genomic signatures. *Cell Rep* 2020;**33**:108296.
94. Tang Z, Kang B, Li C, *et al*. GEPIA2: an enhanced web server for large-scale expression profiling and interactive analysis. *Nucleic Acids Res* 2019;**47**:W556–60.
95. Lauria M, Moyseos P, Priami C. SCUDO: a tool for signature-based clustering of expression profiles. *Nucleic Acids Res* 2015;**43**:W188–92.
96. Wang P, Qin J, Qin Y, *et al*. ChIP-Array 2: integrating multiple omics data to construct gene regulatory networks. *Nucleic Acids Res* 2015;**43**:W264–9.
97. Zhang Y, Yu Z, Ban R, *et al*. DeAnnCNV: a tool for online detection and annotation of copy number variations from whole-exome sequencing data. *Nucleic Acids Res* 2015;**43**:W289–94.
98. Dereeper A, Homa F, Andres G, *et al*. SNiPlay3: a web-based application for exploration and large scale analyses of genomic variations. *Nucleic Acids Res* 2015;**43**: W295–300.
99. Holtgrewe M, Stolpe O, Nieminen M, *et al*. VarFish: comprehensive DNA variant analysis for diagnostics and research. *Nucleic Acids Res* 2020;**48**:W162–9.
100. Jia W, Li H, Li S, *et al*. Oviz-bio: a web-based platform for interactive cancer genomics data visualization. *Nucleic Acids Res* 2020;**48**:W415–26.
101. Sedova M, Iyer M, Li Z, *et al*. Cancer3D 2.0: interactive analysis of 3D patterns of cancer mutations in cancer subsets. *Nucleic Acids Res* 2019;**47**:D895–9.
102. Deng M, Brägelmann J, Kryukov I, *et al*. FirebrowseR: an R client to the broad Institute's firehose pipeline. *Database* 2017;**2017**:baw160.
103. Chauvel C, Novoloaca A, Veyre P, *et al*. Evaluation of integrative clustering methods for the analysis of multi-omics data. *Brief Bioinform* 2020;**21**:541–52.

104. Tini G, Marchetti L, Priami C, *et al*. Multi-omics integration-a comparison of unsupervised clustering methodologies. *Brief Bioinform* 2018;**20**:1269–79.

105. Rappoport N, Shamir R. Multi-omic and multi-view clustering algorithms: review and cancer benchmark. *Nucleic Acids Res* 2018;**46**:10546–62.

106. Eyal-Altman N, Last M, Rubin E. PCM-SABRE: a platform for benchmarking and comparing outcome prediction methods in precision cancer medicine. *BMC Bioinformatics* 2017;**18**.

107. Nguyen H, Shrestha S, Draghici S, *et al*. PINSPlus: a tool for tumor subtype discovery in integrated genomic data. *Bioinformatics* 2019;**35**:2843–6.

108. Dimitrakopoulos C, Hindupur SK, Hafliger L, *et al*. Network-based integration of multi-omics data for prioritizing cancer genes. *Bioinformatics* 2018;**34**:2441–8.