Role of Bioinformatics in Cancer Diagnosis

Jainam H. Valand
Makerere University
Davis Twine (davejtwine@gmail.com)
Makerere University https://orcid.org/0000-0001-8057-9953
Moreen Kyomukamaa
Makerere University
Rebecca Atino
Makerere University
Grace Manana Buteme
Makerere University
Samson Muhahiria
Makerere University
Racheal Nalwoga
Makerere University
Iddy Omary
Makerere University
Anita Grace Nabwami
Makerere University
Emmanuel Otim
Makerere University
David Kabasa
Makerere University
Adam Luyima
Makerere University

Systematic Review

Keywords: cancer, bioinformatics tools, diagnosis, bioinformatics database, gene oncology

Posted Date: January 26th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1299906/v1

Abstract

Cancer is one of the leading causes of mortality around the world accounting for about 10 million deaths in 2020 according to the World Health Organization. The cancer types that claim the most lives around the world include breast cancer, lung cancer, stomach cancer, colon, and rectum cancer. There are a variety of risk factors that can lead to cancers ranging from the type of diet to the type of virus infection. The number of lives claimed by cancer every year can however be reduced through early detection of cancer during which there is a very high chance that the cancer can be cured if appropriate treatment is provided.

Today, due to the development of microarray technology, large amounts of data on differentially expressed genes can be obtained from cancerous cells. This vast amount of data, therefore, requires the use of computational tools and databases to store, process, and extract valuable information from the collected data for example discovering new biomarkers for cancer diagnosis. This, therefore, calls for the application of bioinformatics resources to perform this task. The research article, therefore, focuses on how the different bioinformatics tools and databases have been used to improve cancer diagnosis through a systematic literature search on PubMed. From the literature search, it was seen that bioinformatics tools and databases have been used to detect different diagnostic biomarkers that were associated with the different cancer types such as cervical cancer, ovarian cancer, pancreatic cancer, and lung cancer. The biomarkers detected thus help to improve early cancer detection and hence reduce cancer-related mortality.

From the literature studied, it was also seen that some of the biomarkers detected for one type of cancer were also common to other cancer types. Bioinformatics, therefore, plays a vital role in the improvement of cancer diagnosis by detecting biomarkers that can be used to diagnose cancer. Bioinformatics also helps in identifying common biomarkers and differentially expressed genes in different cancer types which further improves the process of cancer diagnosis.

Background

Cancer is the uncontrolled or unregulated growth of abnormal (malignant or tumor) cells anywhere in the body arising from cells of a specific organ. Cancer arises from the loss of normal growth control of cells and these cells have the ability to create their own blood supply, breaking away from the organ of origin as well as traveling and spreading to other organs of the body(1). Cancer is a genetic disease hence can be inherited or sporadic. Cancer is caused by agents like chemical carcinogens causing DNA mutations, periodic injury, ionizing radiation such as ultraviolet radiations, hormones that stimulate uncontrollable cell growth, genetic abnormalities, immunological dysfunction, viruses like human papillomavirus, hepatitis B, and hepatitis C. (2).

Cancer has become a very big threat to humanity due to its fast growth rate and genomics and remains a frequently lethal disease in humans(the second most frequent cause of death) despite significant

progress made in its diagnosis(3). There are different types of cancer and these include esophageal, oral, bladder, colon, ovarian, lung, breast, gastric, pancreatic, lymphoma, leukemia, glioma, prostate, testicular, melanoma, and hepatoma cancer. Pancreatic cancer is a common malignant tumor of the digestive tract which has a high degree of malignancy and poor prognosis(4). Cervical cancer is one of the highest occurring gynecological cancer that is attributed to high sexual activity with multiple partners, infrequent condom use, and immunosuppression(5).

Cancer diagnosis involves both computational and non-computational diagnosis methods. The non-computational methods include imaging where malignancy is suspected based on imaging information (structural and anatomic). The malignancy is later confirmed on histology and the use of imaging tools permits functional, biochemical, and physiologic assessment of the important aspects of malignancy. Sites for imaging include breasts, brain, lung, and mediastinum using imaging modalities such as Computed Tomography (CT), Magnetic Resonance Imaging (MRI), Ultrasound, Positron Emission Tomography (PET), and Magnetic Resonance Spectroscopy (MRS).

Other non-computational methods include Fluorescence In Situ Hybridization (Fish) technique, tumor markers cytologic and histopathological techniques which involves special staining procedures that can differentiate the different types of tumors e.g., toluidine blue stain. This differentiates mast cell tumors from other tumors as it stains the metachromatic granules present in mast cells. Serological methods such as Enzyme-Linked Immunosorbent Assay (ELISA) and Radioimmuno Assay (RIA) are used in the estimation of serum tumor markers. Immunohistochemistry can also be performed through the use of polyclonal and monoclonal antibodies to detect specific antigenic determinants present in the cells of the tissues. Polymerase Chain Reaction is also being used to establish a definitive diagnosis and classification of tumors based on the recognition of complex profiles that occur in specific tumor types (2).

These non-computational methods, however, have challenges associated with them. One of the challenges faced while using these non-computational methods is low sensitivity and specificity. Sensitivity of a method in the case of cancer diagnosis implies the ability of a method to be able to correctly identify all people in a population with cancer while specificity implies the ability of a method or test to correctly identify all the people in a population without cancer. The sensitivity and specificity values for most non-computational cancer screening methods is in the range of 70–80% and 60–70% respectively(6).

In line with the low sensitivity and specificity of the non-computational methods being used, the Positive Predictive Value (PPV) of the current tumor biomarkers used is also very low which is has led to the failure of cancer screening tests(6). For example, Pap smear, a cervical screening method has sensitivity, specificity, and PPV values of 55.5%, 75%, and 88.2% respectively which are low hence requiring a biopsy in most cases that is highly invasive (7). Due to the low sensitivity of these non-computational methods, there is always a need to perform multiple tests which are quite costly. For example, in a research carried out to evaluate the cost of breast cancer screening in the United States of America, it was seen that about

410 million dollars are spent by women above the age of 75 years in the process of breast cancer screening (8). This thus necessitates the development of genomic bioinformatics technologies to uncover blood-based tumor markers which can greatly improve the specificity and thus boosting the accuracy of the cancer screening process (6).

Currently, microarray technology is widely used in cancer research, diagnosis, and tumor classification for more than a decade. It has been extensively adopted due to limitations with conventional techniques of gene investigation in cancer which are mainly time-consuming and cost-ineffective. Microarrays are significantly advancing due to their small size and are thus applicable when surveying a large number of genes quickly or when the study sample is small. One of the earliest applications of microarray was to identify differences in gene expression between normal and cancer cells. DNA microarray analysis involves the use of an oligonucleotide chip, cDNA chip, and genomic chip. Oligonucleotide microarrays are used for studying gene expression, Single Nucleotide Polymorphism (SNP), mutation, and genotyping analyses, and cDNA microarrays are usually used for gene expression analysis (9).

Microarray technology is a powerful platform for biological exploration. They permit simultaneous analysis of hundreds to thousands of DNA expression sequences for genomic research and diagnostic applications and this provides a guarantee of revolutionizing the way gene expression is examined. In addition to monitoring and analysis of gene expression patterns, microarrays are broadly used to understand the genetic and epigenetic makeup of cancer cells. They are also used to decipher signal pathways of cancer-relevant transcription factors which have advanced the scientific understanding of how cancer-relevant transcription factors control gene networks and ultimately cancer development. Microarray technology also allows the cell's status to be investigated on the molecular scale and can identify a given cell species by its gene expression profile. This is very pivotal in future cancer diagnosis as traditional methods cannot distinguish between morphologically similar but molecularly different tumors. The molecular differences significantly affect the clinical course of a disease (9).

Although the microarray technique for tumor diagnosis and classification is promising as a future diagnostic modality. However, there are many limitations which include the inability to accurately diagnose individual tumors by gene expression profile alone due to the lack of development of a specific group of biomarkers for the diagnosis of specific tumors. This is also attributed to the fact that data analysis tools and methods also need to be developed. In addition, the cost of microarray experiments is high and there are remarkable variations within the same tumor shown by the gene expression data and also early tumor detection is not possible by the gene expression profile. Despite these limitations, DNA microarray is best used for molecular classification based on genetic and biological changes. In conclusion, microarrays are a major tool for the investigation of global gene expression for all aspects of human disease and biomedical research (H. Kim, 2004).

To analyze the vast amounts of data generated by the microarray technology, computational methods are required to be incorporated hence the need for bioinformatics approaches. Bioinformatics is an interdisciplinary approach that integrates information technology with biological science and involves the

creation of databases, the development of software, and data handling for interpretation and analysis on a large scale. The aims of bioinformatics include; first organizing data allowing researchers to easily access existing data as well as submit data into the database. Secondly, it can be used to develop tools and resources that aid in data analysis and finally use the tools to analyze and interpret data in a meaningful biological manner (11). Bioinformatics has applications in several fields such as agriculture, medical science, forensic science, pharmaceutical, and biotech industry (12).

Focusing on the medical applications particularly cancer which is one of the main diseases that destroy lives all over the world, this field of bioinformatics has rapidly grown keeping its pace with the genome sequence expansion (13). Bioinformatics tools/technologies such as web technology, Cytoscape, Gene Expression Profiling Interactive Analysis(GEPIA) and databases such as National Center for Biotechnology Information (NCBI), gene omnibus databases, Surveillance, Epidemiology, and End Results (SEER) database, Kyoto Encyclopedia of Genes and Genomes (KEGG) are being used in cancer research and diagnosis in the identification of biomarkers by analyzing the entire gene expression profiles to approach the disease at a genome level. It is applied in the diagnosis of cancers such as cervical cancer, pancreatic cancer, breast cancer, lung cancer, and several other types. The advancement in bioinformatics technology has thus resulted in faster diagnosis, identification, and prevention of cancer, hence a sustainable solution has been revolutionized with this technology (13).

Main Text

Methodology.

The literature review for the research article involved searching for publications on the PubMed database up to November 2016. To search for the publications, three keywords were used to obtain the publications that involved the applications of bioinformatics in cancer diagnosis. Using the advanced search option on PubMed, the keywords “cancer, “diagnosis”, and “bioinformatics” were first searched on the database using the AND Boolean operator. The search term was, therefore “cancer” AND “diagnosis” AND “bioinformatics” with no terms added as the query leading to the output of 19,798 publications. The search was then modified by adding a query term to search for the keywords in the title as well as the abstract hence the query, title/abstract was used. This modification significantly reduced the search results to 1,080 publications.

The publications were then further filtered to include only research and systematic reviews that further reduced the publications to 75. Publications were then selected and their abstracts reviewed to analyze the relevance of the publication. Publications that did not provide how bioinformatics tools and databases have been used in the diagnosis of cancer were excluded. Articles that were not available in full text and free of cost were also excluded. For the publications that were labeled as relevant, additional information and publications were obtained from the reference sections of the articles. Papers including applications of bioinformatics in the treatment of cancer as well as drug discovery against cancer were also excluded since they were beyond the scope of the topic of study.

Post identifying the papers of relevance and interest, the search results were further filtered by searching for the keywords in only the Title. This, therefore, resulted in the output of 7 publications that were all relevant to the topic of study. For purposes of contrast, publications on the use of non-bioinformatics methods of cancer diagnosis were also obtained. To obtain these publications, both the PubMed database and Google scholar were used. The search for these articles was performed by searching for the keywords “Cancer” and “diagnosis” together with using the AND Boolean operator in the title and abstract of the articles as well as in the title only. This resulted in the output of 11,752 and 1,071 publications respectively. Articles describing relevant information on how cancer is diagnosed using non-bioinformatics tools and databases were then included in the list of articles to be reviewed. The papers were then distributed amongst the 12 authors that reviewed the articles and extracted key information from the articles.

From each of the papers, information such as the findings, the methods used, and key discoveries made were collected and analyzed. The information collected from the articles involving the application of bioinformatics in cancer diagnosis was divided into the bioinformatics databases used for cancer diagnosis, the bioinformatics tools, and software used to analyze the data obtained from the different databases, and the key findings. For the papers related to cancer diagnosis using non-bioinformatics methods, information on the methods of cancer diagnosis and the level of accuracy of the methods were identified.

Bioinformatics tools and databases used in cancer diagnosis

To apply bioinformatics tools and analysis techniques, it is important to first obtain relevant data in line with the area of study. The process of obtaining this data is known as data mining. Data mining involves the use of refined data analysis tools to find unknowns, patterns, and relationships in large data sets. This plays an important role in processes such as gene finding, protein function domain detection, function motif detection, protein function inference, disease diagnosis, disease prognosis, protein and gene interaction network reconstruction, data cleansing, and protein subcellular location prediction (14). A platform that is used in the process of data mining is Oncomine. Oncomine is a cancer microarray database and integrated mining platform that systematically curates analyses and makes available all public cancer microarray data (15).

Bioinformatics databases containing data on cancer

1. The Gene Omnibus Database (GEO).

The GEO database is a public source that archives and distributes high-throughput gene expression and other functional genomics data internationally free of charge. With the rapid change in technologies, the GEO also evolves to expand and include some other data applications like examining chromatin structure and genome-protein interactions rather than only gene expression studies(16). The database has been able to provide access to data for tens of thousands of studies and has also been able to provide various

web-based tools to analyze the data. GEO enables users to visualize and analyze data within their specific interests while providing detailed descriptions(17). The GEO homepage is at http://www.ncbi.nlm.nih.gov/geo/

2. The Cancer Genome Atlas (TCGA).

The Cancer Genome Atlas is one of the most ambitious and successful cancer genomics programs. The database program has generated, analyzed, and made available genomic sequence, expression, methylation, and copy number variation data on over 11,000 individuals representing over 30 different cancer types (18). The Cancer Genome Atlas (TCGA) was a joint effort of the National Cancer Institute (NCI) and the National Human Genome Research Institute (NHGRI), which are both part of the National Institutes of Health, U.S. Department of Health and Human Services (18). The Cancer Genome Atlas, therefore, is a comprehensive and coordinated effort to accelerate our understanding of the molecular basis of cancer through the application of genome analysis technologies, including large-scale genome sequencing.

3. The Human Protein Atlas

The Human Protein Atlas is a Swedish-based program that started in 2003 to map all the human proteins in cells, tissues, and organs using an integration of various omics technologies, including antibody-based imaging, mass spectrometry-based proteomics, transcriptomics, and systems biology. The Human Protein Atlas is divided into three sub-atlases which includes the Tissue Atlas, Cell Atlas, and Pathology Atlas (19). The program relies on sensitive and highly specific antibodies to provide an accurate estimation of protein expression. All antibodies used by the Human Protein Atlas undergo a rigorous validation process which includes Western blot analysis, immunohistochemical staining, and immunofluorescence evaluation against carefully selected sample materials.

The Tissue Atlas contains information regarding the expression profiles of human genes within different tissues at both the protein and mRNA levels. The protein expression data relies on immunohistochemical analysis of 76 different cell types, corresponding to 44 normal human tissue types. The mRNA expression data on the other hand is derived from deep sequencing of RNA from 37 normal tissue types (19).

The Cell Atlas provides information about the spatial distribution of proteins within a panel of 64 cell lines, selected to represent various cell populations in different organs of the human body. The protein expression data is generated by immunofluorescence microscopy, while the mRNA expression data is derived from the deep sequencing of RNA. A subset of these cell lines is subjected to a deeper investigation, within which subcellular protein distribution is classified into 33 different organelles and fine cellular structures.

The Pathology Atlas provides information on how protein expression differs between normal and cancer tissues. The Pathology Atlas contains protein expression and mRNA data for the most common forms of human cancer. The Human Protein Atlas has correlated mRNA expression levels of human genes in cancer tissue with the clinical outcome. The Pathology Atlas, therefore, enables researchers to study

protein expression levels for individual tumors of each cancer type. For example, PAX8 staining is elevated in the thyroid, ovarian, and endometrial cancer tissues.

Post completion of the process of data mining from the databases described in the previous section, bioinformatics tools and techniques are required to perform analysis on the collected data. It is during the analysis process that new components such as biomarkers and biochemical pathways can be discovered to improve the process of cancer diagnosis. Some of the bioinformatics tools that are used to perform the process of data analysis have been described in the following section on bioinformatics tools.

Bioinformatics Tools

1. The Database for Annotation, Visualization and Integrated Discovery (DAVID)

The functional bioinformatics tool uses a collection of algorithms to compress a large group of genes with associated biological terms into relatable well-ordered families, known as biological modules(20). The database is widely being used for complex biological exercises such as but not limited to identifying enhanced biological themes, particularly GO terms; finding out enriched functional-related gene groups; clustering redundant annotation terms; exhibiting connected many-genes-to-many-terms on a 2-D outlook; listing interacting proteins; connecting gene-disease associations and envisaging genes using Bio Carta & KEGG pathway maps.

It mainly uses four data analysis modules which include; the annotation tool that automatically comments on gene lists; GO charts that give visual representations of genes according to the biological process together with molecular function and cellular components; Domain Charts that show the distribution of differentially expressed genes among protein family domains and KEGG Charts through which the differentially expressed genes among KEGG biochemical pathways are exhibited using the KEGG's DBGET, an integrated data retrieval system (21). With over 4,000 journal articles search results on the usage of the DAVID tool in bioinformatics papers in PubMed, over 2,000 studies were cancer-related, showing how the tool has revolutionized the cancer field through assessment of microarray data.

2. Surveillance, Epidemiology, and End Results Program (SEER)

This program launched on January 1 in 1973, aims at gathering information on the diagnosis, treatment, and trends of cancer for about 30% of the United States (U.S) population (22). The program keeps track of the various types of cancer and survival variance by age, ethnicity, and stage at diagnosis time. The program has turned cancer data into discoveries, with over a thousand researchers, clinicians, and legislators using it to analyze and interpret the variations and evolution of cancer in the U.S (NCI, 2018). The SEER program has proven a great tool in observing histopathologic cancer subtypes, and data by molecular subtyping. Over 13,068 studies have actively used the SEER data within the SEER database between 2000 and 2021 with the help of the SEER*Stat software. This has aided in making queries to the

SEER data. Bioinformatics studies have used the SEER database and software to analyze and assess early deaths, survival rates, survival prognostic factors, observe cancer patterns and improve overall outcomes.

3. Gene ontology (GO)

This is a comprehensive bioinformatics resource that provides information about functional genomics to represent biological knowledge. It is a community-based project that is available on (http://www.geneontology.org). The biological knowledge is described in three ways i.e., Molecular function, cell component, and biological process. Molecular function describes the activities performed by gene products and occur at a molecular level like transport and catalysis. The GO rather than describing the complex structures where the activities take place, it provides information on the activities of the gene products.

Cell components provides information on the locations where the gene products perform their activities which may be either the cellular compartment or stable macromolecular complexes. This is in other words the cellular anatomy. Biological processes are the bigger extent processes achieved by several molecular activities. For example, glucose transmembrane transport(23). A ‘GO annotation’ defines the connection of a class from ontology and a gene product with references to the evidence supporting the connection. A Gene Ontology Consortium (GOC) is responsible for monitoring the gene products so that they have consistent descriptions across the biological databases and gene functions across all organisms. The GO can also be used in conjunction with the KEGG path like in (24) that explored this combination to analyze the cancer-related long non-coding RNAs.

4. Gene Expression Profiling Iterative Analysis (GEPIA).

This is a webserver that allows biologists and clinicians to perform comprehensive and complex data mining tasks with simple clicking thus facilitating mining of data in research areas, scientific discussions, and therapeutic discovery about cancer. It is a webserver for profiling and analyzing cancer and normal gene expression (25). In other words, GEPIA provides a tool for resolving bulk RNA datasets in the TCGA and Genotype-Tissue Expression (GTEx) projects to investigate expression profiles across cancer and healthy patient groups. This is done using different techniques like studying its cell- type, interrogating the characteristics of different cell types in cancer(26). It provides a deeper understanding of gene functions and creates new opportunities for data mining in the cancer field of study. (27) used GEPIA to study ovarian cancer expression and prognosis using sirtuins, which are enzymes that have distinct roles in ovarian cancer, and analysis of prognostic biomarkers of cervical cancer done by (28). The webserver is available on http://gepia.cancer-pku.cn/.

5. University of Alabama Cancer Database (UALCAN)

UALCAN database is a comprehensive, user-friendly, and interactive web resource for analyzing cancer omics data. It is an integrated data-mining platform to facilitate the comprehensive analysis of cancer transcriptome. UALCAN uses TCGA RNA-sequencing and patients' clinical data from 33 different cancer types and also includes several metastatic tumors. The web-based platform's user-friendly feature.

UALCAN facilitates relative expression analysis of a query gene(s) across tumor and normal samples. It also identification of the top over- and under-expressed genes in individual cancer types (29).

UALCAN makes it possible to explore or validate the pan-cancer expression pattern of hundreds of user-defined genes. It, therefore, serves as a one-stop-shop by providing easy access to external resources such as Gene Cards, Human Protein Reference Database, PubMed, Target Scan, and Human Protein Atlas that are used to investigate protein expression in various cancers. UALCAN is designed to provide easy access to publicly available cancer OMICS data, allow users to identify biomarkers or perform in silico validation of potential genes of interest and provide graphs and plots depicting expression profile and patient survival information. It is also used to perform cancer gene expression analysis and provide additional information about the selected genes targets by linking to HPRD, Gene Cards, PubMed, Target Scan, and the human protein atlas (29).

Case studies on the application of bioinformatics tools and databases in the diagnosis of cancer

The described bioinformatics tools and databases have been key in improving the process of diagnosis of different cancer types. To understand how the bioinformatics tools and databases described have been used to improve the process of cancer diagnosis, three different case studies were looked at. The case studies looked at involved improving cancer diagnosis of three of the most common cancer types which included cervical cancer, breast cancer, and pancreatic cancer.

a) Cervical cancer; is one of the most dangerous diseases affecting women of all ages. As of 2018, there were approximately 569,000 new cases of cervical cancer worldwide with one of the highest numbers of cases occurring in Uganda (30) and about 311,000 deaths associated with the disease. Of these deaths, about 84-90% of them occurred in low and middle-income countries such as South Africa(31). However, through the combination of high throughput sequencing technology and bioinformatics tools to analyze the data generated, several new gene characteristics and signal pathways that can be used to diagnose cervical cancer have been discovered. The gene characteristics and signal pathways can thus be used to detect the cell pathologies at a very early stage thus improving the process of disease diagnosis, prognosis, and recurrence (32). A study was therefore conducted by Hua-Ju Yang et.al. on how bioinformatics tools and databases can be used to identify key genes and pathways of diagnosis and prognosis in cervical cancer (33).
In the study, three different gene expression profiles consisting of both normal and tumor samples of the same gene were obtained from the Gene Omnibus Database. Genes expressed differently in the normal and tumor cells were then identified using the GEO2R web tool. This process thus provides the starting point of analysis since only the differentially expressed genes would be focused on. The DAVID tool was then used to identify the functional genes and biological pathways from the differentially expressed genes. The Gene Ontology together with the KEGG tools were then used to identify the gene functions, understand biological processes, and also metabolic pathways of the genes (34).

Post identification of the significant genes and their functions, using GEPIA, an interactive web application, the genes were visualized in a box plot format to obtain in-depth information on the genes. Protein-protein interaction was also performed to analyze commonly expressed genes amongst the three gene profiles and the proteins common to two or more genes. Based on the different analyses performed using the different bioinformatics tools, 12 key differentially expressed genes were discovered from a total of 57 differentially expressed genes that were involved in processes such as cell division and epithelial cell differentiation. All the genes identified had a high level of expression in cervical cancer tissues compared to normal tissues (33).

A key gene found and identified as CXCL8 was common to all cervical cancer tissues and was also associated with poor prognosis in patients with high expression of the gene. This was also verified from UALCAN online tool using data from the Oncomine database. The gene was also associated with other cancer types such as pancreatic cancer, head and neck tumors, breast cancer, and many others. The analysis performed also found that the gene plays a key role in apoptosis resistance and tumorigenesis. Cervical cancer patients with decreased levels of the CXCL8 gene have also been shown to have a better survival rate. Other key genes identified included the MCM2, TOP2A, TYMS, and HELLS genes. These genes were identified to be responsible for the occurrence of cervical cancer and thus were considered to be vital diagnostic markers (33). The bioinformatics tools and database have thus been used to identify key differentially expressed genes which can be used as biomarkers to detect cancer at a very early stage and improve the patient survival chances.

b) Breast cancer; Breast cancer represents a top biomedical research issue since it's the most frequently diagnosed cancer in females with increasing mortality rates in past years. Early diagnosis and treatment of breast cancer is of paramount importance. Breast cancer diagnosis is done using Magnetic resonance imaging, ultrasound, mammography, positron emission tomography, and biopsy, but these methods are; expensive, with low sensitivity, and lengthy. Bioinformatics presents a quicker and more efficient approach by identification of breast cancer biomarkers for early detection and hence the diagnosis of breast cancer (35).
DEGs in breast cancer we established using 3 datasets of the GEO database. Analysis of genome pathways was used to show the functional roles of differentially expressed genes. The Chinese breast cancer tissues by (Reverse Transcriptase – quantitative Polymerase Chain Reaction) RT-qPCR were used to authenticate expression of novel DEGs with a total of 46. Two novel biomarkers; ADH1A and IGSF10, and 4 other genes (APOD, KIT, RBP4, and SFRP1) were seen as causes of breast cancer since they were expressed in breast cancer tissues. Also, 14 out of 25 microRNAs targeting 6 genes were seen to be associated with breast cancer and hence potential biomarkers for diagnosis of breast cancer (32).

Weighted gene co-expression analysis was done with Gene Set Enrichment Analysis (GSEA) for genome-wide RNA expression. Functional enrichment analysis, GO was used to describe the function of gene and gene products. KEGG analysis was used to annotate genes with pathway and functional information. Breast cancer cell lines were obtained from the Pathology Laboratory of the Cancer Institute of the Fourth

Hospital of Hebei Medical University. Cell lines were maintained in DMEM-H medium supplemented with fetal bovine serum(35).

c) Pancreatic cancer; In the digestive tract pancreatic cancer is common with a poor prognosis. Early detection of biomarkers of pancreatic cancer is good for timely detection and management to improve prognosis and lower mortality rates attributed to it. In this study, DLGAP5 expression in pancreatic cancer was explored in tumorigenesis and tumor growth. So, differentially expressed genes were screened via the GEO data set GSE16515. GO based functional analysis & KEGG pathways enrichment analysis was performed on the conforming proteins of the genes using the DAVID. Analysis was done using the Kaplan–Meier Plotter database to establish the relationship between differentially expressed genes and pancreatic cancer prognosis. The DLGAP5 gene was isolated and its expression in pancreatic cancer and other cancer tissues was profiled using the Oncomine and GEPIA databases (4).
The overall DLGAP5 survival was analyzed using the TCGA database. Thereafter molecular mechanisms of pancreatic cancer were analyzed by GSEA. Finally, a cell function experiment was done to discover the DLGAP5 biological behavior. During this study, 201 upregulated differentially expressed genes & 79 downregulated genes were used. And tumor-related signaling pathways were observed with emphasis on; the cancer pathways, extracellular matrix-receptor interaction pathway & p53 signaling pathway. It was found that the DLGAP5 was highly expressed in pancreatic cancerous cells and the higher the levels the worse the prognosis. This gene had also seriously enriched in cell signaling pathways e.g. the cell cycle, p53, and oocyte meiosis. (4).

Conclusion

In conclusion, cancer being a leading cause of death worldwide remains a very challenging field. Its effective early diagnosis is therefore paramount, which however is very challenging with the non-computational methods that are less accurate, expensive and some like CT scans involves exposure to radiation. Bioinformatics, therefore, has become a very powerful and revolutionizing technique in improving cancer diagnosis. It aids the early cancer diagnosis of the different types of cancers, by discovering key biomarkers using a range of tools and databases, which utilizes the vast amounts of data generated by microarray technology.

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and materials

Not applicable

Competing interests

The authors declare that they have no competing interests.

Funding

Not applicable

Authors' contributions

All authors collected and analyzed information; JHV, DT, MK, RA, GBM edited and revised the manuscript; JHV drafted the manuscript. All the authors read, reviewed, and approved the final manuscript before submission.

Acknowledgments

Technical support in writing the research article was provided by Mr. Julius Mugaga, Department of Physiology, Biomedical Engineering, Makerere University.

References

1. Becker L. Cancer Notes. Leaving Art. 2013;(January):211–21.
2. Kumar P, Pawaiya RVS. Advances in cancer diagnostics. Brazilian J Vet Pathol. 2010;3(2):142–53.
3. Samarasinghe S, Ganegoda GU. Bioinformatics For Cancer Diagnosis. 2020;(March).
4. Ke MJ, Ji LD, Li YX. Bioinformatics analysis combined with experiments to explore potential prognostic factors for pancreatic cancer. Cancer Cell Int. 2020 Aug;20(1):1–13.
5. Mwaka AD, Orach CG, Were EM, Lyratzopoulos G, Wabinga H, Roland M. Awareness of cervical cancer risk factors and symptoms: cross-sectional community survey in post-conflict northern Uganda. Health Expect [Internet]. 2016 Aug 1 [cited 2021 Dec 28];19(4):854. Available from: /pmc/articles/PMC4957614/
6. Schiffman JD, Fisher PG, Gibbs P. Early Detection of Cancer : Past, Present, and Future INTRODUCTION TO CANCER SCREENING AND. ASCO Educ B. 2015;
7. Nkwabong E, Laure Bessi Badjan I, Sando Z. Pap smear accuracy for the diagnosis of cervical precancerous lesions. Trop Doct. 2019;49(1):34–9.

8. Gross CP, Long JB, Ross JS, Abu-Khalaf MM, Wang R, Killelea BK, et al. The cost of breast cancer screening in the Medicare population. JAMA Intern Med [Internet]. 2013 Jan 11 [cited 2021 Dec 28];173(3):220–6. Available from: https://pubmed.ncbi.nlm.nih.gov/23303200/
9. Kim I-J, Kang HC, Park J-G. Microarray Applications in Cancer Research. Cancer Res Treat. 2004;36(4):207.
10. Kim H. Role of M icroarray in C ancer D iagnosis. 2004;36(1):1–3.
11. Luscombe NM, Greenbaum D, Gerstein M. Mim01-Luscombe-What-Is-Bioinf. 2001;346–58.
12. Singh SK. Bioinformatics : Concepts and Applications Chapter 12 Bioinformatics : Concepts and Applications. 2020;(July).
13. Kihara D, Yang YD, Hawkins T. Bioinformatics resources for cancer research with an emphasis on gene function and structure prediction tools. 2006;25–35.
14. Yu S, Guan Y, Dai Y. Application of data mining in ESMC. Jisuanji Gongcheng/Computer Eng. 2003;29(19):90.
15. Rhodes DR, Yu J, Shanker K, Deshpande N, Varambally R, Ghosh D, et al. ONCOMINE: A Cancer Microarray Database and Integrated Data-Mining Platform. Neoplasia. 2004;6(1):1–6.
16. Clough E, Barrett T. The Gene Expression Omnibus database. Methods Mol Biol. 2016;1418:93–110.
17. Barrett T, Suzek TO, Troup DB, Wilhite SE, Ngau WC, Ledoux P, et al. NCBI GEO: Mining millions of expression profiles - Database and tools. Nucleic Acids Res. 2005;33(DATABASE ISS.):562–6.
18. Wang Z, Jensen MA, Zenklusen JC. A Practical Guide to The Cancer Genome Atlas (TCGA). Methods Mol Biol. 2016;1418:111–41.
19. Lindskog C. The Human Protein Atlas – an important resource for basic and clinical research. https://doi.org/101080/14789450201611992 80. 2016 Jul;13(7):627–9.
20. Huang DW, Sherman BT, Tan Q, Collins JR, Alvord WG, Roayaei J, et al. The DAVID Gene Functional Classification Tool: A novel biological module-centric algorithm to functionally analyze large gene lists. Genome Biol. 2007;8(9).
21. Dennis G, Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, et al. DAVID: Database for Annotation, Visualization, and Integrated Discovery. Genome Biol. 2003;4(5).
22. Duggan MA, Anderson WF, Altekruse S, Penberthy L, Sherman ME. The Surveillance, Epidemiology, and End Results (SEER) Program and Pathology. Am J Surg Pathol. 2016;40(12):e94–102.
23. Blake JA, Christie KR, Dolan ME, Drabkin HJ, Hill DP, Ni L, et al. Gene ontology consortium: Going forward. Nucleic Acids Res. 2015;43(D1):D1049–56.
24. Chen L, Zhang YH, Lu G, Huang T, Cai YD. Analysis of cancer-related lncRNAs using gene ontology and KEGG pathways. Artif Intell Med. 2017;76:27–36.
25. Tang Z, Li C, Kang B, Gao G, Li C, Zhang Z. GEPIA: A web server for cancer and normal gene expression profiling and interactive analyses. Nucleic Acids Res. 2017;45(W1):W98–102.
26. Li C, Tang Z, Zhang W, Ye Z, Liu F. GEPIA2021: Integrating multiple deconvolution-based analysis into GEPIA. Nucleic Acids Res. 2021;49(W1):W242–6.

27. Sun X, Wang S, Li Q. Comprehensive Analysis of Expression and Prognostic Value of Sirtuins in Ovarian Cancer. Front Genet. 2019;10(September):1–14.
28. Liu J, Liu S, Yang X. Construction of Gene Modules and Analysis of Prognostic Biomarkers for Cervical Cancer by Weighted Gene Co-Expression Network Analysis. Front Oncol. 2021;11(March):1–13.
29. Chandrashekar DS, Bashel B, Balasubramanya SAH, Creighton CJ, Ponce-Rodriguez I, Chakravarthi BVSK, et al. UALCAN: A Portal for Facilitating Tumor Subgroup Gene Expression and Survival Analyses. Neoplasia. 2017 Aug;19(8):649–58.
30. Nakisige C, Trawin J, Mitchell-Foster S, Payne BA, Rawat A, Mithani N, et al. Integrated cervical cancer screening in Mayuge District Uganda (ASPIRE Mayuge): A pragmatic sequential cluster randomized trial protocol. BMC Public Health. 2020 Jan 31;20(1).
31. Hull1 R, Mbele M, Makhafola T, Hicks C, Wang SM, Reis RM, et al. Cervical cancer in low and middle-income countries. Oncol Lett [Internet]. 2020 Sep 1 [cited 2021 Dec 5];20(3):2058–74. Available from: https://pubmed.ncbi.nlm.nih.gov/32782524/
32. Martínez-Rodríguez F, Limones-González JE, Mendoza-Almanza B, Esparza-Ibarra EL, Gallegos-Flores PI, Ayala-Luján JL, et al. Understanding Cervical Cancer through Proteomics. Cells [Internet]. 2021 Aug 1 [cited 2021 Dec 5];10(8). Available from: https://pubmed.ncbi.nlm.nih.gov/34440623/
33. Yang H ju, Xue J min, Li J, Wan L hong, Zhu Y xi. Identification of key genes and pathways of diagnosis and prognosis in cervical cancer by bioinformatics analysis. Mol Genet genomic Med [Internet]. 2020 Jun 1 [cited 2021 Dec 5];8(6). Available from: https://pubmed.ncbi.nlm.nih.gov/32181600/
34. Khaket TP, Aggarwal H, Dhanda S, Singh J. Enzyme informatics. Curr Top Med Chem [Internet]. 2012 Jul 1 [cited 2021 Dec 5];12(17):110–43. Available from: https://pubmed.ncbi.nlm.nih.gov/23116471/
35. Shan Y. The Role of PAX2 in Breast Cancer: A Study Based on Bioinformatics Analysis and in Vitro Validation. 2021 Jul;