International Journal of Recent Trends in Engineering & Research
ISSN(ONLINE) : 2455-1457
IMPACT FACTOR : 3.344

Development of a NGS Cancer Research Database – CancerBase

Quashiya M. Soudagar[1], Akshatha Prasanna[2], V. G. Shanmuga Priya[3]
[1]M.Tech, Bioinformatics, KLE Dr. M.S Sheshgiri College of Engineering and Technology, Belagavi
[2]Head R & D, GenEclat Technologies, Bengaluru
[3]Asst. Professor, Dept. of Biotechnology, KLE Dr. M.S Sheshgiri College of Engineering and Technology

***A*bstract:** Cancer is one of the most common causes of death all over the world. According to the World Health Organization, it causes about 12.5% (or 12.5 out of every 100) of all deaths worldwide. It is caused due to the mutations in the human DNA over a period of time which results in an abnormal growth of cells. There is a need to design a database with the description of specific cancer features which helps the medical users and researchers for the better understanding of the cancer mechanism to provide better diagnostics and treatment. Our study aims at the development of a cancer specific database i.e. CancerBase which provides information on Next Generation Sequencing SRA information and mutations causing cancer and the pathways involved.
Keywords: Cancer; Mutations; Genes; CancerBase; NGS-SRA

I. INTRODUCTION

Cancer is a type of disease in which an abnormal growth of cells occurs which divide indivisibly and invade other tissues and cells. Usually, the healthy cells divide in a controlled manner and copy themselves to create new healthy cells without any defect. But, in a cancer patient, this process of normal cell division goes abnormal and the cell division takes place without control. The cells change their nature as the mutations occur in their respective genes. And the daughter cells of the cancer cells are also cancerous. Cancer can affect anybody of any age. But, it is most prominent in older people as their immunity decreases and the DNA may get damaged or altered. There are more than 100 types of cancer, including breast cancer, bladder cancer, kidney cancer, pancreatic cancer, skin cancer, lung cancer etc. There may be different symptoms based on the types of the cancers. The various symptoms may be acute cough, fever, abnormal bleeding, lump, sudden weight gain or loss, change in bowel movements etc. The causes may be chemicals, dyes, heredity, infection, radiation, hormones, physical agents etc. The main causes are mutations in the genes. Hence, a database is created as a repository for this information. The database contains the NGS-SRA based data. The database is created using the PHP for front end and MySQL for back end. WampServer 2.5 is used for the Connections to the database. The Next-Generation Sequencing (NGS), with its unprecedented throughput, scalability, and speed, next-generation sequencing enables researchers to study biological systems at a level never possible before. The SRA is The Sequence Read Archive, previously known as the Short Read Archive is a public repository for DNA sequencing data, especially the "short reads" generated by High-throughput sequencing, which are typically less than 1,000 base pairs in length.

A database management system (**DBMS**) is a computer software application that interacts with the user, other applications, and the database itself to capture and analyze data. A biological database is a large, organized body of persistent data, usually associated with computerized software designed to update, query, and retrieve components of the data stored within the system. MySQL is structured query language used as the backend for the database. It is an open source relational database management

system and is widely used client-server model. It is a special-purpose programming language designed for managing data held in a relational database management system.

To implement the queries and build a website, PHP (Hypertext Preprocessor) is used at the front end. PHP is a server-side scripting language developed for the websites and is also used as a general-purpose programming language. It was earlier known as the Personal Home Page. The PHP code can be embedded into HTML code and can also be used in combination with various web template systems, web frameworks etc.

II. MATERIALS AND METHODS

The database contains the information such as Cancer type, SRR ID, Source of sample, Strategy of experiment, Layout, mutated Gene list, Gene Description, GO Id, Pathways etc. In order to achieve these goals, researchers must be able to effectively store, access, and manipulate the huge amount of short read data generated from massively parallel sequencing experiments.

The tools used are:
Programming languages: PHP
Database language: MySQL
Server: WampServer 2.5
Databases used for Analysis: SRA, GeneCard, SNPedia.

WampServer is a software stack for the Microsoft Windows operating system, developed by Romain Bourdon and consisting of the Apache web server, OpenSSL for SSL support, MySQL database and PHP programming language. The WampServer 2.5 version is used in this work as it is compatible with the available operating system.

The NGS-SRA data for various cancers are selected with additional information on strategy, layout and sample source from NCBI-SRA. The mutations causing the cancer are studied from various literatures and cross validated from available databases. The top deregulated genes in the cancer are added to our database for the respective cancer type selected. SNP causing the cancer disease is retrieved from SNPedia and clinvar. The respective mutation sequence is provided in the text format. The pathways affected by the mutations are found out and are updated to the database.

A database is created in the Wampserver 2.5 and the information is retrieved using the MySQL query language. PHP code is embedded in HTML code and the connection is established with the database. Now, the website is created with a user-friendly graphical interface.

A. WORKFLOW:

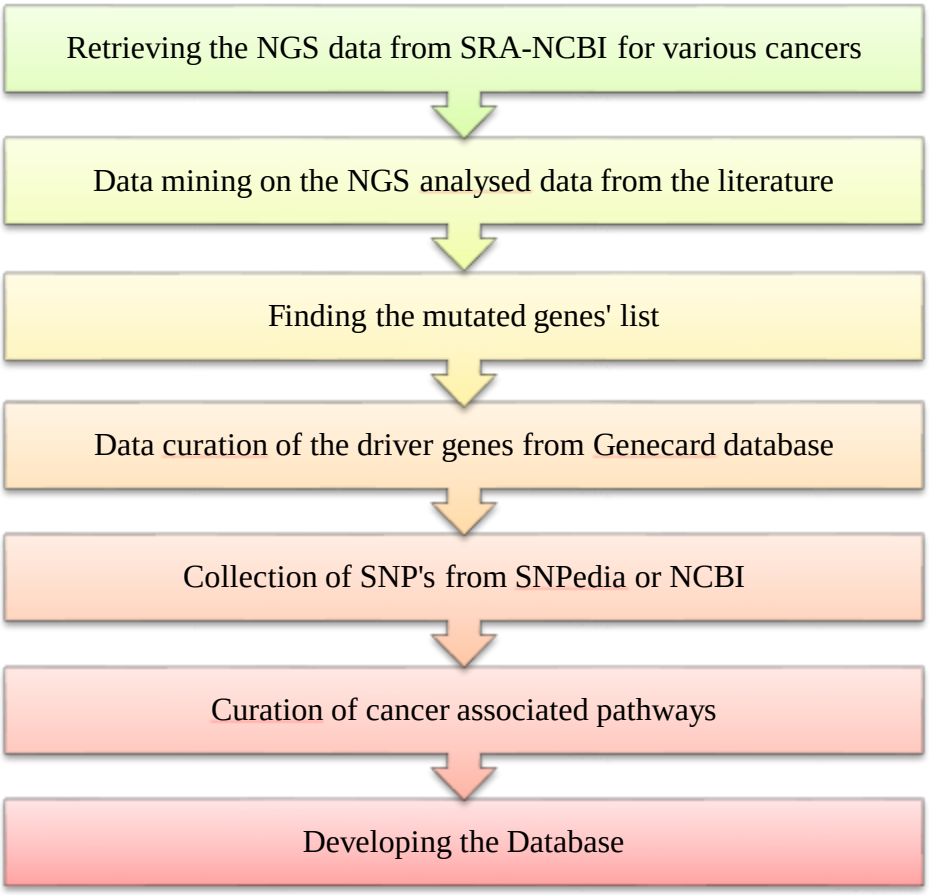

Figure 1. Flowchart of the process

III. RESULTS

The results are as follows which depict the information on SRA data and the genes.

Strategy- Sequencing strategy used in the experiment
Source- Type of genetic source material sequenced
Layout- Configuration of the read layout may be Single, Paired, Fragment, etc.

For example, for Breast cancer, the SRA Sequence Id is ERR1307002, the strategy is miRNA-Seq, Source is TRANSCRIPTOMIC and the Layout is SINGLE.

Cancer type	SRA Sequence ID	Strategy	Source	Layout
Bladder cancer	SRR3056095	OTHER	GENOMIC	PAIRED
Brain tumour	SRR522150	RNA-Seq	TRANSCRIPTOMIC	SINGLE
Breast cancer	ERR1307002	miRNA-Seq	TRANSCRIPTOMIC	SINGLE
Cervical cancer	SRX1537778	AMPLICON	GENOMIC	SINGLE

Figure 2. SRA data

Cancer Type	GENE	Gene Description	GO ID	Pathways
Bladder cancer	TP53	Tumor Protein P53	GO:0000981	CDK-mediated phosphorylation and removal of Cdc6
	ERBB2	Erb-B2 Receptor Tyrosine Kinase 2	GO:0001042	Interleukin receptor SHC signaling
	EGFR	Epidermal Growth Factor Receptor	GO:0003682	Apoptotic Pathways in Synovial Fibroblasts
	CDH1	Cadherin 1, Type 1	GO:0001948	Signaling by Rho GTPases
	KRAS	Kirsten Rat Sarcoma Viral Oncogene Homolog	GO:0005515	IL-7 Signaling Pathways
Brain tumour	TP53	Tumor Protein P53	GO:0000981	CDK-mediated phosphorylation and removal of Cdc6
	TNF	Tumor Necrosis Factor	GO:0005164	PEDF Induced Signaling
	GFAP	Glial Fibrillary Acidic Protein	GO:0005178	Interleukin receptor SHC signaling
	IL6	Interleukin 6	GO:0005138	Apoptotic Pathways in Synovial Fibroblast
	PTEN	Phosphatase And Tensin Homolog	GO:0004722	PI3K/AKT Signaling in Cancer
Breast cancer	TP53	Tumor Protein P53	GO:0000981	CDK-mediated phosphorylation and removal of Cdc6
	BRCA1	Breast Cancer 1, Early Onset	GO:0003677	DNA Double-Strand Break Repair
	BRCA2	Breast Cancer 2, Early Onset	GO:0003697	DNA Double-Strand Break Repair
	ERBB2	Erb-B2 Receptor Tyrosine Kinase 2	GO:0001042	PI3K/AKT Signaling in Cancer
	ESR1	Estrogen Receptor 1	GO:0001046	Nuclear Receptor transcription pathway
	TP53	Tumor Protein P53	GO:0000981	Glioma

Figure 3. Gene Information

A database website is created using MySQL and PHP with the following graphical user interface.

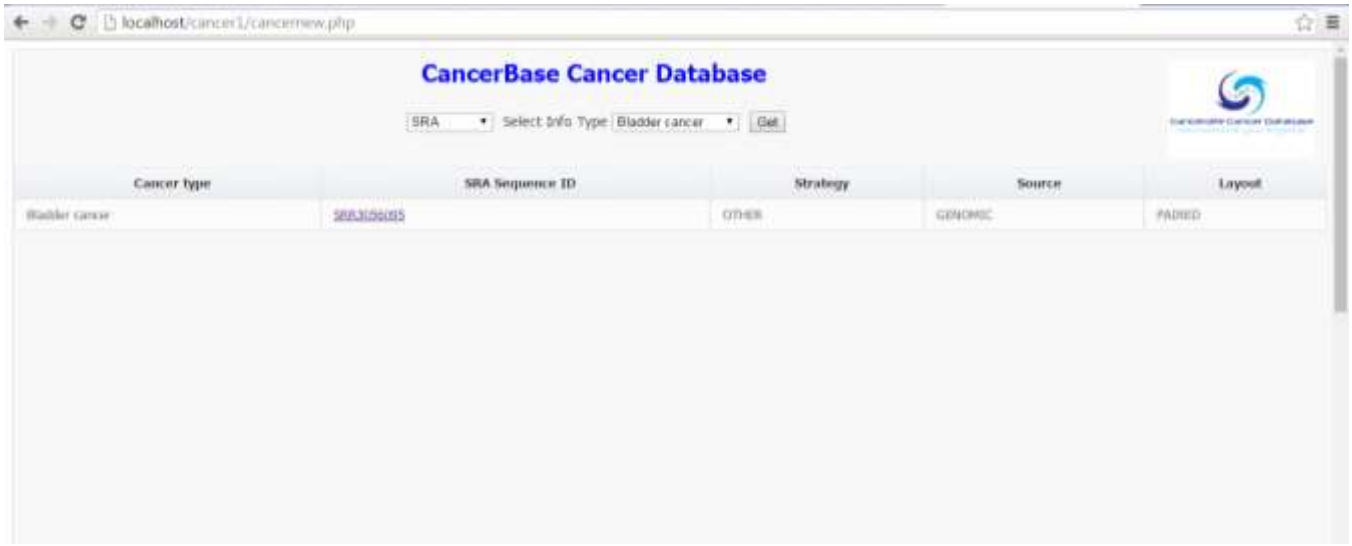

Figure 4. The CancerBase Database for SRA information

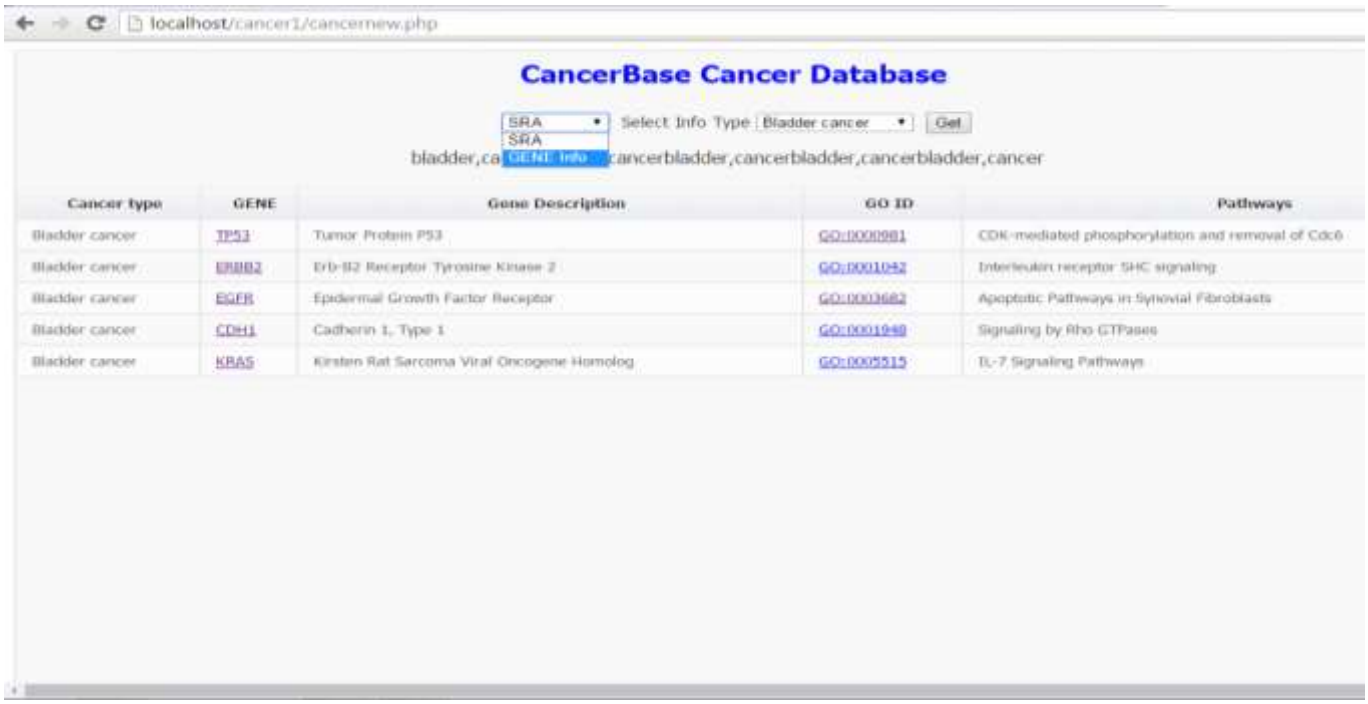

Figure 4. The CancerBase Database for gene information

ACKNOWLEDGEMENTS

I would like to take this opportunity to give my sincere and heartfelt thanks to our beloved Principal Dr. Basavaraj G. Katageri, K.L.E Dr. M S Sheshgiri College of Engineering and Technology, Belagavi for providing me with all the facilities in the college. Then I would like to thank our HOD, Dr. S. C. Mali, Department of Biotechnology for his constant support. I also extend my thanks to my guide Ms. Akshatha Prasanna and Mr. Naveen Kumar, GenEclat Technologies, Bengaluru. Then, let me show my sincere gratitude to my internal guide Asst. Prof. V. G. Shanmuga Priya, Dept. of Biotechnology, for her persistent guidance. Again I am obliged to my parents and friends for giving support in the entire endeavour for the completion of my project. Above all I am grateful to the Almighty, without his help I could not accomplish anything.

REFERENCES

1. Forbes SA, Tang G, Bindal N, Bamford S, Dawson E, Cole C, Kok CY, Jia M, Ewing R, Menzies A, et al. COSMIC (the Catalogue of Somatic Mutations in Cancer): a resource to investigate acquired mutations in human cancer. Nucleic Acids Res. 2010;38:D652–D657.
2. Haider S, Ballester B, Smedley D, Zhang J, Rice P, Kasprzyk A. BioMart central portal – unified access to biological data. Nucleic Acids Res. 2009;37:W23–W27.
3. Petitjean A, Mathe E, Kato S, Ishioka C, Tavtigian SV, Hainaut P, Olivier M. Impact of mutant p53 functional properties on TP53 mutation patterns and tumor phenotype: lessons from recent developments in the IARC TP53 database. Hum. Mutat. 2007;28:622–629.
4. Sjöblom T, Jones S, Wood LD, Parsons DW, Lin J, Barber TD, Mandelker D, Leary RJ, Ptak J, Silliman N, et al. The consensus coding sequences of human breast and colorectal cancers. Science. 2006;314:268–274. [4]. Tomasetti M, Neuzil J, Dong L. MicroRNAs as regulators of mitochondrial function: role in cancer suppression. Biochimica et Biophysica Acta 2014;1840(4):1441-1453.
5. Stein LD, Mungall C, Shu S, Caudy M, Mangone M, Day A, Nickerson E, Stajich JE, Harris TW, Arva A, et al. The generic genome browser: a building block for a model organism system database. Genome Res. 2002;12:1599–1610.

