A Review of Online Tools for Cancer Genomics Studies

Jatinder Garg[1], Sonu Bala Garg[2*]

[1]Baba Hira Singh Bhattal Institute of Engineering and Technology, Lehragaga (Punjab)
[2]I.K.Gujral Punjab Technical University, Jalandhar, Punjab, India
**Corresponding Author: Email: sonugarg79@yahoo.com,*

ABSTRACT

Bioinformatics has emerged out as a branch of computer science which is concerned with helping biologists in managing, analysing, interpreting, storing and sharing data. Studies and research in the field cancer generates large amount of such data. A variety of computational tools and databases have been developed worldwide to help cancer scientist in data management and sharing. In this paper, an attempt was made to survey such tool available at various locations on the internet and presenting them in tabular form with a brief introduction about each, for ready reference of cancer scientists.

Keywords: Bioinformatics, Cancer, Genomics, Data Integration

1. INTRODUCTION

Developing computation skills, to help researchers and scientists in diverse fields of science, is the prime objective of computer scientists today. Research in medical science generate large amount of data that is needed to be analysed and managed to draw meaningful conclusions. Bioinformatics emerged out as a branch of science in which biology, computer science and information technology are blended to develop computations tools for understanding, analysis and interpreting biological data [1]. It also involves developing new algorithms and statistical measures to establish relationship among various members of large biological data sets [2].

In twenty first century cancer has emerged out to be one of the deadliestdiseases in the world in general and India in particular. It has been found that cancer is basically a disease of genome [1]. Cancer genomics is the study of the genetic abnormalities that leads to cancer. It primarily deals with finding new cancer genes, new therapeutic targets and to identify molecular signatures for stratifying tumors [Ref]. Recent advancements in cancer study has led to the gain of a lot of knowledge which is primarily in the form of data about complete genome sequences including point mutation and structure alterations for various types of cancer [1]. The usefulness of this type of data will depend upon the availability of computational tools for effectively interpreting such data. A variety of such tools and algorithms have been developed by scientists and technologists throughout the world, but these tools are scattered and discretely available without any coordination to make them available at a central place. As a result of this some of the very fine algorithms remain largely unused or underused.

This paper is an attempt to explore and summaries some of the very best bioinformatics tools available in the field of cancer research. A brief introduction along with their availability information has been provided for each tool. Although the list is not exhaustive, but it will definitely help the medical researchers working in the field of cancer to select the best tools for getting optimal results.

2. VARIOUS TOOLS FOR CANCER GENOMICS STUDY

Various online tools available for cancer genomic studies have been summerised and presented in the form of Table 1 given below:-

Table 1: Various online tools available for cancer genomic studies

Tool	URL	Features
methBLAST and methPrimerDB	http://medgen.ugent.be/methprimerdb http://medgen.ugent.be/methblast	Modification of critical genes, Sequence alterations, Search portal
ICGC	https://dcc.icgc.org	Multidisciplinary information, Data analysis, Enhancement of accuracy of diagnosis
CMS	http://cbbiweb.uthscsa.edu/KMethylomes	Comprehensive collection of datasets, Visualization and analysis of methylation datasets for cancer
DREMECELS	http://www.bioinfoindia.org/dremecels	Relationship among mirnas, drug sensitivity, somatic mutations and methylation
MENT	http://mgrc.kribb.re.kr:8080/MENT	Correlation between DNA methylation and gene expression, Visualization
TCGA	http:// genomeportal.stanford.edu/pan-tcga	Correlation between TCGA genomic results and clinical phenotype, Searching for clinical parameters
cBio Cancer Genomics Portal	http://cbioportal.org	Exploration of multidimensional cancer genomics data sets, Clinical applications, Visualization, Somatic mutations
canEvolve	http://www.canevolve.org	Copy number alterations, Storage and visualization of clinical genes co-expression, Correlation between gene expression and clinical outcomes.
Genomics and Public Health	http://www.cdc.gov/pcd/issues/2005	To help practitioners increase their awareness of the impact of genomics
canSAR	http://cansar.icr.ac.uk	Cancer translational research and drug discovery, Multidisciplinary annotations for biological systems and genes, Chemical screening, RNA interference screening, Easy access to multidisciplinary data
NONCODE	http://www.bioinfo.org/noncode	Collection and annotation of non-coding RNAS, Relationship between lncrnas and diseases, RNA sequencing
Cancer Genomics Hub	https://cghub.ucsc.edu	Largest repository of data, Data storage and transmission
EGA	https://www.ebi.ac.uk/ega	Distribution and sharing of genetic and phenotypic data, Data storage, Security

MethyCancer	http://methycancer.psych.ac.cn	Data analysis, Alteration of DNAmethylation, Correlation between DNA methylation, Gene expression, Mutation and cancer
Integrative Genomics Viewer	http://www.broadinstitute.org/igv	Expression profiling of coding and non codingRNAS, Visualization, Sequence reads, Mutations, Copy number, RNAI screens, Gene expression, Methylation, and genomic annotations
Pancreatic expression database	http://www.pancreasexpression.org	Largest collection of multidimensional pancreatic data, Generate increased volume of data, Robust and rigorous data mining
Cascade	https://github.com/aaronshifman/Cascade_RNAseq_viewer	Rna-seq, Visualization, Cancer genomics, Dimensionality reduction

The details are as under:-

A. methBLAST and methPrimerDB

DNA methylation plays an important role in development and tumorigenesis by epigenetic modification and silencing of critical genes. Bisulphite DNA modification results in sequence alterations which results in the development of the methBLAST, a sequence similarity search program based on the original BLAST algorithm. A part from the specific analysis tool, a public database is also developed named methPrimerDB for the storage and retrieval of validated PCR based methylation assays. Database records can be searched by gene symbol, nucleotide sequence, analytical method used, Entrez Gene or methPrimerDB identifier and submitter's name. PCR based methylation analysis methods to study human, mouse and rat epigenetic modifications [2].

B. International Cancer Genome Consortium Data Portal

The International Cancer Genome Consortium (ICGC) is a multidisciplinary, multi-institutional collaborative effort to characterize somatic mutations in 50 different cancer types and sub- types. It contains data from 24 cancer projects, including ICGC, The Cancer Genome Atlas (TCGA), Johns Hopkins University, and the Tumor Sequencing Project. It consists of 3478 genomes and 13 cancer types and subtypes. ICGC data types include simple somatic mutations, copy number alterations, structural rearrangements, gene expression, microRNAs, DNA methylation and exon junctions. The Data Portal uses a web-based graphical user interface (GUI) to offer researchers multiple ways to quickly and easily search and analyze the available data. The major goals of ICGC is to rapidly bring these data to cancer research community in order to accelerate studies on the discovery of cancer causes to enhance the accuracy of diagnoses and to improve treatments [3].

C. Cancer Methylome System

Cancer Methylome System (CMS) is designed for the visualization, comparison and statistical analysis of human cancer-specific DNA methylation. Methylation intensities were obtained from MBDCap-sequencing, pre-processed and stored in the database. 191 patient samples (169 tumor and 22 normal specimen) and 41 breast cancer cell-lines are deposited in the database, comprising about 6.6 billion uniquely mapped sequence reads. CMS includes important analytic functions for interpretation of methylation data, such as the detection of differentially methylated regions, statistical calculation of global methylation intensities, multiple gene sets of biologically significant categories, interactivity with UCSC via custom-track data. CMS can be visualized in two distinct modes: genomic view and gene centric view. CMS is freely accessible at http:// cbbiweb.uthscsa.edu/KMethylomes/ [4].

D. DREMECELS

It mainly provides qualitative and quantitative information on these cancer types along with methylation, drug sensitivity, miRNAs, copy number variation (CNV) and somatic mutations data. The database contains 156 genes and two repair mechanisms, base excision repair (BER) and mismatch repair (MMR). The database comprises of diversified data, such as basic gene and protein information, functional annotation, literature references, associated transcription factors, conserved domains information, gene ontology (GO) information, miRNA information, interacting partners, path- ways, methylation, drug details, copy number variation (CNV), somatic mutations, and pro- vides suitable links to various primary external resources. DREMECELS is publicly available at http://www.bioinfoindia.org/dremecels [5].

E. MENT

The objective of MENT is to provide researchers information on both DNA methylation and gene expression in diverse cancers. It contains integrated data of DNA methylation, gene expression, correlation of DNA methylation and gene expression in paired samples and clinicopathological conditions gathered from TCGA and GEO. MENTS is the first database which provides both DNA methylation and gene expression information in diverse normal and tumor tissues. Two aspects of MENT are unique. First, the overall boxplot summary of methylation and gene expression patterns in diverse normal and tumor tissues provides users an invaluable. instant insight on the gene of interest. Second, providing both DNA methylation and gene expression data allows users to understand the effects of DNA methylation on gene expression [6].

F. TCGA

The Cancer Genome Atlas (TCGA) project has generated genomic, epigenomic, transcriptomic, and proteomic data for over 20 different cancer types. These data files included genomic, transcriptomic, epigenomic, and proteomic data for each of the 25 cancer types. these data included DNA CNV, somatic mutations, mRNA expression level by RNA sequencing (RNA-Seq), DNA methylation, miR expression level by RNA-Seq, and protein expression level by RPPA. The Cancer Genome Atlas Clinical Explorer is a clinically oriented summary of genomic/proteomic data organized by cancer type or clinical parameters. Its interface enables users to query TCGA data in multiple ways. This profiling function requires three inputs including a gene/miR/protein, a cancer type, and a clinical parameter. Cancer Genome Atlas Clinical Explorer is accessible at http://genomeportal.stanford.edu/pan-tcga [7].

G. cBio Cancer Genomics Portal

The cBio Cancer Genomics Portal is an open-access resource for interactive exploration of multidimensional cancer genomics data sets, currently providing access to data from more than 5,000 tumor samples from 20 cancer studies. The cBio Cancer Genomics Portal significantly lowers the barriers between complex genomic data and cancer researchers who want rapid, intuitive, and high-quality access to molecular profiles and clinical attributes from large- scale cancer genomics projects and empowers researchers to translate these rich data sets into biologic insights and clinical applications. The cBio portal currently contains 5 published data sets (2–5) and 15 provisional TCGA data sets. In addition to mutation data, the portal includes copy number alterations, microarray-based and RNA sequencing–based mRNA expression changes, DNA methylation values, and protein and phosphor protein levels [8].

H. canEvolve

canEvolve query functionalities are designed to fulfill most frequent analysis needs of cancer researchers with a view to generate novel hypotheses. canEvolve stores gene, microRNA (miRNA) and protein expression profiles, copy number alterations for multiple cancer types, and protein-protein interaction information. At present canEvolve provides different types of information extracted from 90 cancer genomics studies comprising of more than 10,000 patients. canEvolve portal to help cancer biologists easily access the knowledge and analysis results derived from primary, integrative and network analysis of oncogenomic data generated using various functional genomics platforms. The focus of this work is the generation of the database framework capable of storing multiple data types and the user-friendly web interface. It also allows visualization of regulatory and protein- protein interaction networks [9].

I. Genomics and Public Health

In 2001, the Centers for Disease Control and Prevention funded three Centers for Genomics and Public Health to develop training tools for increasing genomic awareness. have developed tools to increase awareness of the impact genomics will have on public health practice. During the past few years, several genomics training tools have been developed to help public health practitioners increase their awareness of the impact of genomics on public health practice, such as genetic testing for adult cancer, identifying genetically at-risk subgroups susceptible to environmental exposures, or developing new genomic technologies. These training tools aim to provide a foundation for understanding basic genomics (e.g., DNA mutations, inheritance patterns). They also help practitioners identify and translate the relevance of genomics to their own work (e.g., using family history as a genomic–environmental indicator of a person’s own risk of chronic diseases) [10].

J. canSAR

canSAR is a fully integrated cancer research and drug discovery resource developed to utilize the growing publicly available biological annotation, chemical screening, RNA interference screening, expression, amplification and 3D structural data. This allows easy access to the multidisciplinary data within, including target and compound synopses, bioactivity views and expert tools for chemo genomic, expression and protein interaction network data. canSAR integrates genomic, protein, pharmacological, drug and chemical data with structural biology, protein networks and druggability data. canSAR is widely used to rapidly access information and help interpret experimental data in a translational and drug discovery context. It is used by >150 000 unique users from 179 countries, and is used by biologists, chemists and translational and clinical scientists, from both academia and industry. canSAR contains the full complement of the human proteome as well as 528 805 proteins from 16 634 model organisms and data for 11 778 cancer and non-transformed cell line models [11].

K. NONCODE

NONCODE is an interactive database that aims to present the most complete collection and annotation of non-coding RNAs, especially long non-coding RNAs (lncRNAs). The recently reduced cost of RNA sequencing has produced an explosion of newly identified data. NONCODE contains 527,336 lncRNA transcripts from 16 species (human, mouse, cow, rat, chimpanzee, etc.). NON- CODE 2016 has also introduced three important new features: (i) conservation annotation; (ii) the relation- ships between lncRNAs and diseases; and (iii) an interface to choose high-quality datasets through predicted scores, literature support and long-read sequencing method support [12].

L. Cancer Genomics Hub (CGHub)

Cancer genomic hub, which was established in August 2011, is the largest warehouse of cancer genomic data. It has been funded and managed by National Cancer Institute, Maryland and is hosted by University of California. The database has grown to the size of around 2.5 petabytes at present and is serving a download of around 3 byte every month. The data downloading is as simple as if being downloaded from a hard disk. These data specifically are designed for three main components that are managed by the National Cancer Institute (NCI), including The Cancer Genome Atlas (TCGA), the Cancer Cell Line Encyclopedia (CCLE) and the Therapeutically Applicable Research to Generate Effective Treatments (TARGET) project, among others [13].

M. European Genome Phenome Archive

The European Genomephenome Archive (EGA) is a permanent archive which is used for the distribution and sharing of genetic and phenotypic data. The EGA follows strict protocols for information management, data storage, security and dissemination. The European Genome phenome Archive (EGA) is a permanent repository for all types of potentially identifiable genetic and phenotypic data from biomedical research projects. The EGA provides secure access to restricted data for authorized researchers and clinicians. EGA provide information on how access decisions are made, outline the methods for data submission and dissemination, and describe the EGA system infrastructure [14].

N. MethyCancer

DNA methylation plays a critical role in tumorigenesis through regulating oncogene activation, tumor suppressor gene silencing and chromosomal instability. For DNA methylation, gene expression and cancer, publicly accessible database is developed for human DNA Methylation and Cancer (MethyCancer, http:// methycancer.genomics.org.cn). MethyCancer hosts both highly integrated data of DNA methylation, cancer-related gene, mutation and cancer. There are mainly four types of data included in MethyCancer: (i) CGI clones and global CGI predictions, (ii) DNA methylation data, (iii) cancer information, genes and mutations and (iv) correlation among DNA methylation, gene expression and cancer [15].

O. Integrative Genomics Viewer

Rapid improvements in sequencing and array-based platforms are resulting in a flood of diverse genome-wide data, including data from exome and whole genome sequencing, epigenetic surveys, expression profiling of coding and non-coding RNAs, SNP and copy number profiling, and functional assays. (IGV), a lightweight visualization tool that enables intuitive real-time exploration of diverse large- scale genomic datasets, on standard desktop computers. It supports flexible integration of a wide range of genomic data types including aligned sequence reads, mutations, copy number, RNAi screens, gene expression, methylation, and genomic annotations. IGV supports concurrent visualization of diverse data types across hundreds, and up to thousands of samples, and correlation of these integrated datasets with clinical and phenotypic variables [16].

P. Pancreatic expression database

The Pancreatic Expression Database is the only device currently available for mining of pancreatic cancer literature data. It brings together the largest collection of multidimensional pancreatic data from the literature including genomic, proteomic, microRNA, methylomic and transcriptomic profiles. We added a user guide and implemented integrated graphical tools to overlay and visualize retrieved information. Interoperability with biomart-compatible data sets was significantly improved to allow integrative queries with pancreatic cancer data. PED data are divided into two separate data sets. The ‘Gene Expression Data Set’ contains relevant data from transcriptomic, proteomic, methylomic and miRNA studies, and the Copy Number Alteration Data Set’ contains copy number alteration (CNA) data from genomic studies [17].

Q. Cascade

Cascade, a novel web-based tool for the intuitive 3D visualization of RNA-seq data from cancer genomics experiments. The Cascade viewer allows multiple data types (e.g. mutation, gene expression, alternative splicing frequency) to be simultaneously displayed, allowing a simplified view of the data in a way that is tunable based on user specified parameters. Cascade, a new data visualization tool to display and explore NGS datasets in a rapid and intuitive way by allowing multiple data attributes to be shown simultaneously. Cascade allows the analysis of RNA-seq data, or whole-exome or genome sequencing, to be easily mapped onto known or user defined biological path- ways. Cascade is a web-based user interface that allows re- searchers to interactively explore their RNA-seq data while allowing a wide variety of data types to be displayed. Cascade consists of the main web page, an underlying relational database (MySQL) containing all of the information from RNA-Seq experiments (along with information defined in biological pathways, gene lists, etc.) and a collection of PHP scripts allowing the user to submit requests to the database to be displayed in the browser [18].

3. CONCLUSION

For the interpretation and analysis of large amount of biological data generated in cancer research, a variety of computational resources have been developed globally. Majority of such tool are available online and can be accessed freely. Some of such prominent resources were studied, compiled and presented at a central place to help the cancer scientists pick up the most appropriate tool concerned with their research. Resources such as MENT, Methy Cancer are used for correlation between DNA methylation and Gene expression, ICGC, canSar, Methy cancer for data analysis & accuracy of diagnosis, CG hub, EGA, methBLAST and meth primer db for data storage, CMS for Visualisation, Dremecels, Cascade, NONCODE, IGV for RNA sequencing, TCGA, Cbioportal for clinical parameters.

REFERENCES

[1] Yadong Yang, Xunong Dong, BingbingXie, Nan Ding, Juan Chen, Yongjun Li, Qian Zhang, Hongzhu Qu, Xiangdong Fang, "Databases and Web Tools for Cancer Genomics Study", Genomics Proteomics Bioinformatics, 2015.

[2] Filip Pattyn, JasmeinHoebeeck, Piet Robbrecht, EviMichels, Anne De Paepe, et.al, "methBLAST and methPrimerDB: web - tools for PCR based methylation analysis", BMC Bioinformatics, 2006.

[3] Junjun Zhang, Joachim Baran, A. Cros, Jonathan M. Guberman, Syed Haider, Jack Hsu, Yong Liang, Elena Rivkin, Jianxin Wang, Brett Whitty, Marie Wong-Erasmus, Long Yao and ArekKasprzyk, "International Cancer Genome Consortium Data Portal—a one-stop shop for cancer genomics data".

[4] Fei Gu, Mark S. Doderer, Yi-Wen Huang, Juan C. Roa, Paul J. Goodfellow, E. Lynette Kizer, Tim H. M. Huang, Yidong Chen, "CMS: A Web-Based System for Visualization and Analysis of Genome-Wide Methylation Data of Human Cancers", PLOS ONE, April 2013.

[5] Ankita Shukla, Ahmed Moussa, Tiratha Raj Singh, "DREMECELS: A Curated Database for Base Excision and Mismatch Repair Mechanisms Associated Human Malignancies", PLoS ONE, June 2016.

[6] Su-JinBaek, Sungjin yang, Tae-Wook Kang, Seong-Min Park, Yong Sung Kim, Seon-Young Kim, "MENT: Methylation and expression database of normal and tumor tissues".

[7] HoJoon Lee, Jennifer Palm, Susan M. Grimes and Hanlee P. Ji, "The Cancer Genome Atlas Clinical Explorer: a web and mobile interface for identifying clinical–genomic driver associations", Genome Medicine (2015).

[8] Ethan Cerami, Jianjiong Gao, UgurDogrusoz, Benjamin E. Gross, SelcukOnur Sumer, Bülent Arman Aksoy, Anders Jacobsen et.al, "The cBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data", Cancer Discov. 2012.

[9] Mehmet Kemal Samur, Zhenyu Yan, Xujun Wang, Qingyi Cao, Nikhil C. Munshi, Cheng Li, Parantu K. Shah, " canEvolve: A Web Portal for Integrative Oncogenomics", PLOS ONE, February 2013.

[10] Jennifer Bodzin, MPH, Sharon L.R. Kardia, Aaron Goldenberg, Sarah F. Raup, Janice V. Bach, Toby Citrin, "Genomics and Public Health: Development of Web-based Training Tools for Increasing Genomic Awareness", Centers for Disease Control and Prevention, April 2005.

[11] Joseph E. Tym, Costas Mitsopoulos, Elizabeth A. Coker, Parisa Razaz, Amanda C. Schierz, Albert A. Antolin and Bissan Al-Lazikani, "canSAR: an updated cancer research and drug discovery knowledgebase", Nucleic Acids Research, 2016.

[12] Yi Zhao1, Hui Li, Shuangsang Fang, Yue Kang, Wei wu, Yajing Hao, Ziyang Li, Dechao Bu, Ninghui Sun, Michael Q. Zhang, and Runsheng Chen, "NONCODE 2016: an informative and valuable data source of long non-coding RNAs", Nucleic Acids Research, 2016.

[13] Christopher Wilks, Melissa S. Cline, Erich Weiler, Mark Diehkans, Brian Craft, Christy Martin, Daniel Murphy, Howdy Pierce, et.al, "The Cancer Genomics Hub (CGHub): overcoming cancer through the power of torrential data ", Database (2014), bau093.

[14] IlkkaLappalainen, Jeff Almeida-King, Vasudev Kumanduri, Alexander Senf, John Dylan Spalding, Saifur-Rehman, Gary Saunders, Jag Kandasamy, Mario Caccamo, "The European Genome-phenome Archive of human data consented for biomedical research", Nature Genetics, July 2015.

[15] Ximiao He, Suhua Chang, Jiajie Zhang, Qian Zhao, Haizhen Xiang, KanthidaKusonmano, Liu

Yang, Zhong Sheng Sun, Huanming Yang and Jing Wang, "MethyCancer: the database of human DNA methylation and cancer ", Nucleic Acids Research, 2008.

[16] James T. Robinson, Helga Thorvaldsdóttir, Wendy Winckler, Mitchell Guttman, Eric S. Lander, Gad Getz, and Jill P. Mesirov, "Integrative Genomics Viewer", Nat Biotechnol. 2011.

[17] Abu Z. Dayem Ullah, Rosalind J. Cutts, MillikaGhetia, EmanuelaGadaleta, Stephan A. Hahn, Tatjana Crnogorac-Jurcevic, Nicholas R. Lemoine and Claude Chelala, " The pancreatic expression database: recent extensions and updates ", Nucleic Acids Research, 2014.

[18] Aaron R. Shifman, Radia M. Johnson and Brian T. Wilhelm, "Cascade: an RNA-seq visualization tool for cancer genomics", BMC Genomics (2016).