Genomics Proteomics Bioinformatics 13 (2015) 46–50

HOSTED BY

ELSEVIER

Genomics Proteomics Bioinformatics

www.elsevier.com/locate/gpb
www.sciencedirect.com

RESOURCE REVIEW

Databases and Web Tools for Cancer Genomics Study

Yadong Yang [1,a], Xunong Dong [1,2,b], Bingbing Xie [1,2,c], Nan Ding [1,2,d], Juan Chen [3,e], Yongjun Li [1,f], Qian Zhang [1,g], Hongzhu Qu [1,h], Xiangdong Fang [1,*,i]

[1] *CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100101, China*
[2] *University of Chinese Academy of Sciences, Beijing 100049, China*
[3] *Technical Service Center of Family Planning and Reproductive Health, National Research Institution for Health and Family Planning, Beijing 100081, China*

Received 30 December 2014; revised 26 January 2015; accepted 27 January 2015
Available online 21 February 2015

Handled by Qiang Tian

KEYWORDS

Cancer;
Genomics;
Data integration;
Resource;
Collaboration

Abstract Publicly-accessible resources have promoted the advance of scientific discovery. The era of genomics and big data has brought the need for collaboration and data sharing in order to make effective use of this new knowledge. Here, we describe the web resources for cancer genomics research and rate them on the basis of the diversity of cancer types, sample size, omics data comprehensiveness, and user experience. The resources reviewed include data repository and analysis tools; and we hope such introduction will promote the awareness and facilitate the usage of these resources in the cancer research community.

Introduction

There has been accumulating evidence over the last two to three decades supporting that cancer is a disease of the genome. Previous studies have followed a one-by-one approach to examine the molecular mechanisms of cancer, although this approach is one-sided and inefficient. With the development of high-throughput sequencing technologies, recent years have witnessed a great data explosion and systematic study of the cancer genome. For the first time, data were made available for the complete genome sequences including point mutations and structural alternations for a large number of cancer types, enabling the differentiation of cancer subtypes in an unprecedented global view. However, the effective use of the massive

* Corresponding author.
E-mail: fangxd@big.ac.cn (Fang X).
[a] ORCID: 0000-0003-2936-1574.
[b] ORCID: 0000-0002-0956-502X.
[c] ORCID: 0000-0002-8573-442X.
[d] ORCID: 0000-0002-1045-1695.
[e] ORCID: 0000-0001-9901-1524.
[f] ORCID: 0000-0002-2122-1721.
[g] ORCID: 0000-0003-4580-171X.
[h] ORCID: 0000-0001-7013-8409.
[i] ORCID: 0000-0002-6628-8620.

Peer review under responsibility of Beijing Institute of Genomics, Chinese Academy of Sciences and Genetics Society of China.

http://dx.doi.org/10.1016/j.gpb.2015.01.005
1672-0229

amounts of cancer genome data remains a challenge due to the limitations of computational methodologies and insufficient collaboration and sharing (**Figure 1**). In this paper, we describe several popular and effective web-based cancer genomics data repositories, along with tools and resources (**Table 1**) to manage and analyze these data. We have rated the resources based on data comprehensiveness and ease-of-use according to our own experience.

Cancer Genomics Hub

The Cancer Genomics Hub (CGHub) is a central repository for the genomic information generated through three different programs at the National Cancer Institute (NCI) of the United States, namely, The Cancer Genome Atlas (TCGA), the Cancer Cell Line Encyclopedia (CCLE), and the Therapeutically Applicable Research to Generate Effective Treatments (TARGET) projects [1]. CGHub is hosted at the University of California, Santa Cruz (UCSC), with controlled data access to ensure patient privacy. CGHub holds nearly 1.9 PB of data, covering 42 cancer types and normal controls (https://cghub.ucsc.edu/summary_stats.html). Till Dec 2014, there have been more than 10,000 samples from TCGA alone (https://tcga-data.nci.nih.gov/tcga/tcgaHome2.jsp) and more than 500 papers have been published by researchers from the TCGA Research Network and those who used TCGA data in their work (http://cancergenome.nih.gov/publications). The launch of CGHub will promote the sharing of cancer data, collaboration between cancer researchers, and potentially facilitate the personalized medicine.

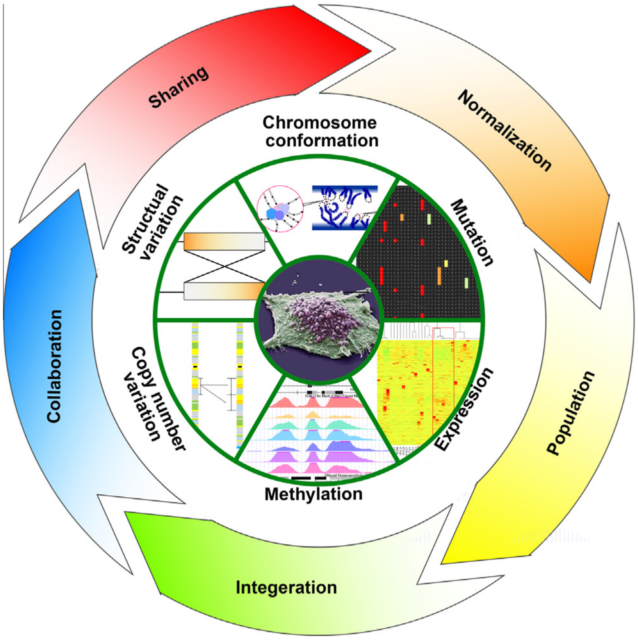

Figure 1 The future of cancer research
In the era of big data, one of the major challenges is to make full use of multi-dimensional data from heterogeneous sources, including different omics data and a variety of medical data from bedside. The success in the battle against cancer will largely depend on population-sized information from both genomic and clinical resources, advanced algorithms for data mining, open and sharing circumstances.

European Genome-phenome Archive

The European Genome-phenome Archive (EGA) is a data center for all types of sequencing and genotyping experiments. Almost 58% of all studies in EGA are related to cancer, including data generated by the International Cancer Genome Consortium (ICGC). Since its founding in 2008, ICGC has produced terabytes of data from about 12,232 donors and 50 cancer projects (https://dcc.icgc.org/repository/release_17). Somatic variant data are openly accessible at the ICGC Data Portal (https://dcc.icgc.org/repository), whereas raw sequence data, germline mutations, and clinical data are held at EGA with controlled access.

Catalogue Of Somatic Mutations In Cancer

The Catalogue Of Somatic Mutations In Cancer (COSMIC) is the largest database of somatic mutations and their effects on human cancer [2]. The database is curated manually from published literature, allowing very precise definitions of disease types and patient details. The database is updated every 2 months and has thus far integrated 15,047 whole cancer genomes from 1,058,292 samples, including 2,710,499 coding mutations, 10,567 gene fusions, 61,232 genomic rearrangements, 702,652 copy number variations (CNVs), and 118,886,698 abnormal expression variants. Data can be queried by key words and downloaded by registered users. COSMIC has also stored curated, large-scale systematic screens and whole-genome shotgun sequencing papers for reference. The huge, manually-curated, and regularly-updated dataset of the COSMIC database makes it an invaluable resource for cancer studies.

Cancer Program Resource Gateway

The Broad Institute is one of the most famous academic centers for cancer study. Its Cancer Program aims to investigate the fundamental mechanisms of cancer and research from discovery to clinical applications. The Program releases many datasets and tools for scientific research, which are held at the Broad Cancer Program Resource Gateway (CPRG). We will describe Broad's Genome Data Analysis Center (GDAC), one of these resources, as an example.

Broad's GDAC

It is important but generally time-consuming or sometimes even impossible for most labs to analyze terabytes of sequence data. However, the Firehose system from Broad's GDAC is changing the status quo. The GDAC systematically analyzes data from TCGA pilot and extends to other diseases as well. Firehose now assembles ~40 terabytes of TCGA data and reliably executes more than 6000 pipelines per month. GDAC obtains and processes the TCGA data every 2 weeks, and makes them available afterward [3]. Firehose contains series of standardized pipelines for genomic analysis and the computing environment is accessible to the public so that people can install and run their own tools for data analysis. Taking advantage of the powerful computational environment at