Sci. Bull. (2015) 60(1):65–70
DOI 10.1007/s11434-014-0692-9

www.scibull.com
www.springer.com/scp

The cancer genomics and global cancer genome collaboration

Xueda Hu · Huanming Yang · Jie He · Youyong Lu

Received: 20 November 2014 / Accepted: 26 November 2014 / Published online: 25 December 2014

Abstract All cancers arise as a result of abnormalities occurring in the DNA sequence of cancer cells, and we are now stepping into an era in which it is feasible to obtain the complete DNA sequence of large cohorts of cancer patients. The International Cancer Genome Consortium (ICGC) launched in 2007 is devoted to coordinate large-scale cancer genome studies in tumors from 50 different cancer types and/or subtypes and systematic studies of more than 25,000 cancer genomes. Several participant groups have summarized and published their data for various cancers. As the active members of ICGC, Chinese cancer genome investigators have contributed research for 13 tumor types and released some research articles about esophageal, liver, bladder, and kidney cancers. As genetic alterations in thousands of tumors have now been catalogued, the pan-cancer analysis has become the most significant role of ICGC at present. The ICGC research network will reveal the repertoire of oncogenic mutations, uncover traces of the mutagenic influences, define molecular subtypes for clinical implication, and enable the development of individual therapeutics for human cancers.

Keywords Cancer genome · International Cancer Genome Consortium · Sequencing · Mutation

1 The origination and current status of International Cancer Genome Consortium

The cellular and molecular mechanisms of malignancy have been studied for more than a century. Numerous hypotheses have been proposed, from chromosome abnormalities to cancer-causing genes (oncogenes and tumor suppressor genes) [1, 2]. Many seminal but important discoveries have validated the search for cancer-related DNA alterations underlying the development of human cancer. In 1986, Renato Dulbecco [3] wrote an influential commentary and suggested that the complete sequence of the human genome would be an essential tool for systematically discovering the genes that drive cancer. Finally, the Human Genome Project (HGP) was launched in 1990, with a "draft" sequence completed by 2000 [4, 5] and a near-complete sequence published in 2004 [6].

The availability of the human genome sequence brings in a great opportunity to systematically analyze DNA changes in a cancer cell, and the research community has rapidly begun to form a new field of "cancer genomics". Pioneering work started sequencing large numbers of PCR products of total or partial protein-coding regions to identify somatic point mutations and small insertions and deletions (Indels). With the development and widespread use of massively parallel sequencing platforms, the full range of genetic alteration in cancer could then be assayed. The National Cancer Institute of the United States and Wellcome Trust

X. Hu · J. He
Department of Thoracic Surgery, Cancer Institute and Hospital, Chinese Academy of Medical Sciences, Beijing 100021, China

H. Yang
BGI-Shenzhen, Shenzhen 518083, China

H. Yang
James D. Watson Institute of Genome Sciences, Hangzhou 310029, China

Y. Lu (✉)
Laboratory of Molecular Oncology, Key Laboratory of Carcinogenesis and Translational Research (Ministry of Education), Peking University Cancer Hospital and Institute, Beijing 100142, China
e-mail: youyonglu@hsc.pku.edu.cn

Sanger Institute in the United Kingdom have launched large-scale cancer genome initiatives, called The Cancer Genome Atlas (TCGA, http://cancergenome.nih.gov/) and the Cancer Genome Project (CGP, http://www.sanger.ac.uk/research/projects/cancergenome/), respectively. The United States has also initiated the Therapeutically Applicable Research to Generate Effective Treatments (TARGET, https://ocg.cancer.gov/programs/target) project specifically to catalogue genetic changes in pediatric tumors. Following these individual efforts, cancer genome scientists and funding agencies met in Toronto, Canada, in October 2007 to discuss and launch an international collaboration called the International Cancer Genome Consortium (ICGC).

2 Coordination of global cancer genome project: the scope and progression of the ICGC

The ICGC is committed to bring together global researchers to comprehensively analyze the genomic, transcriptomic, and epigenomic changes in at least 50 different tumor types or subtypes that are of clinical and societal importance across the globe (http://www.icgc.org). The genetic alterations for each sample recorded should catalogue the full range of somatic changes, including single-nucleotide variants (SNV), Indels, copy number variations (CNV), translocations, and other chromosomal rearrangements. The genomic data should reach the following features as the steering committee of ICGC indicated in their maker paper [7]: (1) comprehensiveness, sequencing a cumulative number of >25,000 tumor genomes and discovering cancer genes with somatic abnormalities occurring at a frequency of >3%; (2) high resolution, at a single-nucleotide level; (3) high quality, using common standards for pathology and technology; (4) data from matched non-tumor tissue, to distinguish somatic from inherited sequence variants and aberrations; and (5) generate complementary catalogues of transcriptome and epigenome datasets from the same tumors. The ICGC will make the data available to entire research community as rapidly as possible, without compromising the need to protect patient confidentiality, in accordance with the Toronto statement [8].

As of April 2014, the ICGC has received commitments from global researchers and funding organizations for 74 project teams in 17 jurisdictions. Over 27,000 tumors that meet ICGC requirements for consent and pathology have been obtained (Fig. 1a). Those specimens are across 22 tumor types in a broad pathological classification (Fig. 1b) and are comprised of 2,520 pediatric tumors and 24,426 adult tumors (Fig. 1c). Considering the low percentage of pediatric cancer samples, the ICGC steering committee is discussing how to promote more pediatric cancer patients and projects to integrate into the ICGC network. Raw datasets exist for out of 18,000 tumors, including 1,096 whole-genome sequences, 5,560 exomes, 16,936 copy number alteration profiles, 17,297 transcriptomes (RNA-Seq, miRNA-Seq, and array-based expression), and 9,017 epigenomes submitted to ICGC. Processed data are available via the Data Coordination Centre (DCC, http://dcc.icgc.org) based at the Ontario Institute for Cancer Research and is updated bimonthly. As of May 2014 (version 16), this includes datasets from 49 ICGC members and comprises data from 11,633 cancer genomes (Fig. 1d). The data released from the ICGC DCC are under surveillance of the Data Access Compliance Office (DACO) of ICGC. Up to May 2014, the DACO had approved 497 data application for a variety non-commercial uses of the ICGC raw datasets.

3 The scientific findings of ICGC participants

Several laboratories participated in ICGC collaboration have already reported scientific outcomes derived from the cancer genome data. Cancer evolves dynamically as cell clones expansion competitively, which is presented by its somatic mutation profile. The Sanger Institute developed a mathematical method to extract mutational signatures of the underlying processes of cancer tissue development and applied it to 21 breast cancers [9, 10]. The Sanger group has also reported novel genomic changes in myelodysplasia. A frequently recurrent somatic SF3B1 mutation was detected in myelodysplasia with ring sideroblasts, implicating abnormalities of messenger RNA splicing in the pathogenesis of myelodysplastic syndromes. Clinically, patients with SF3B1 mutations had fewer cytopenias and longer event-free survival than wild-type patients [11]. The researchers also evaluated the clinical significance of this mutated gene in other myelodysplastic syndromes and myelodysplastic/myeloproliferative neoplasms [12]. They suggest a hypothesis of genetic "predestination" that early driver mutations, typically affecting genes involved in RNA splicing, dictate future trajectories of disease evolution with distinct clinical phenotypes [13]. The Indian ICGC group has been focusing on genomic exploration of oral cancer. Through exome sequencing and recurrence testing, they found USP9X, MLL4, ARID2, UNC13C, and TRPM3 frequently altered in gingivobuccal oral squamous cell carcinoma (OSCC-GB), as well DROSHA, YAP1 with recurrent amplifications, and DDX3X with homozygous deletions in OSCC-GB [14]. From the ICGC PedBrain Tumor Project, researchers have identified recurrent activating mutations in FGFR1 and PTPN11, and new NTRK2 fusion genes in pilocytic astrocytoma [15]. For Ewing sarcoma, French ICGC researchers have found STAG2