REVIEW ARTICLES

The impact of the Cancer Genome Atlas on lung cancer

JEREMY T.-H. CHANG, YEE MING LEE, and R. STEPHANIE HUANG

CHICAGO, ILL

The Cancer Genome Atlas (TCGA) has profiled more than 10,000 samples derived from 33 types of cancer to date, with the goal of improving our understanding of the molecular basis of cancer and advancing our ability to diagnose, treat, and prevent cancer. This review focuses on lung cancer as it is the leading cause of cancer-related mortality worldwide in both men and women. Particularly, non-small cell lung cancers (including lung adenocarcinoma and lung squamous cell carcinoma) were evaluated. Our goal was to demonstrate the impact of TCGA on lung cancer research under 4 themes: diagnostic markers, disease progression markers, novel therapeutic targets, and novel tools. Examples are given related to DNA mutation, copy number variation, messenger RNA, and microRNA expression along with methylation profiling. (Translational Research 2015;166:568–585)

Abbreviations: CIMP = CpG island methylator phenotype; GSEA = gene set enrichment analysis; LUAD = lung adenocarcinoma; LUSC = lung squamous cell carcinoma; miRNA = microRNA; NSCLC = non-small cell lung carcinoma; OS = overall survival; TCGA = the Cancer Genome Atlas

OVERVIEW OF THE CANCER GENOME ATLAS

The Cancer Genome Atlas (TCGA) project is a major collaborative effort being undertaken in the United States to advance our "understanding of the molecular basis of cancer through the application of genome analysis technologies, including large-scale genome sequencing."[1] Because cancer is viewed as a complex genetic disorder, there is sustained interest in finding the genetic pathways and aberrant genomic changes that precipitate in the development of tumors. The ultimate goal of TCGA is to "improve our ability to diagnose, treat, and prevent cancer" through the discoveries and insights enabled by the exhaustive mapping of various forms of cancer.[1]

Originally established by the National Institute of Health, TCGA is jointly run by the National Cancer Institute and the National Human Genome Research Institute. The 3-year pilot project of TCGA began in 2006 to develop the policies and infrastructure to handle the high-volume data generated from patient samples. In 2009, the National Cancer Institute and National Human Genome Research Institute announced that TCGA was ready to move forward into a new 5-year project. This project would comprehensively analyze the genomic landscape of more than 20 cancers, ranging from glioblastomas to prostate adenocarcinomas.[2]

From the Biological Sciences Collegiate Division, The University of Chicago, Chicago, Ill; Center for Personalized Therapeutics, The University of Chicago, Chicago, Ill; Department of Medicine, The University of Chicago, Chicago, Ill.

Submitted for publication June 30, 2015; accepted for publication August 3, 2015.

Reprint requests: R. Stephanie Huang, 900 E 57th street, KCBD Room 7148, The University of Chicago, Chicago, IL 60637; e-mail: rhuang@medicine.bsd.uchicago.edu.

1931-5244/$ - see front matter

http://dx.doi.org/10.1016/j.trsl.2015.08.001

The cancers studied by TCGA were selected for their "poor prognosis and overall public health impact," and the "availability of human tumor and matched normal tissue samples that meet TCGA standards for patient consent, quality, and quantity."[3] The general strategy for this unprecedented effort was to collect large samples of each tumor type along with matched normal tissue, with some tumor types receiving more than 500 unique samples.

To generate comprehensive molecular profiles for each tumor type, different technologies were used by TCGA, such as whole genome and exon sequencing, single nucleotide polymorphism (SNP) genotyping, copy number variation (CNV) profiling using microarrays (ie, Agilent, Illumina), DNA methylation profiling, genome-wide expression, functional proteomic analysis, and microRNA (miRNA) expression profiling through RNA sequencing. This rich amount of information generated from TCGA was then consolidated and shared with the community.

By January 2015, TCGA announced it had successfully collected "the necessary quality and quantity of samples" for all 33 selected tumor types. Table I presents a summary of the available TCGA data for these 33 tumor types. To date, publicly available genomic data have been shared online for 29 of the 33 tumor types. As of 2015, breast invasive carcinoma comprised the highest number of data with 1098 cases, whereas cholangiocarcinoma had the lowest number with 36 cases.[4] TCGA data have spawned more than 2700 articles in scientific journals, demonstrating the far-reaching effects of TCGA data on cancer research.[5] As TCGA winds down, new projects such as PanCanAtlas and Pan-cancer analysis of whole genomes are now underway to further analyze TCGA data and to gain an even clearer understanding of the molecular mechanisms of cancer.[5] This review seeks to focus on lung cancer as it is the leading cause of cancer-related mortality worldwide in both men and women. The review aims to demonstrate the impact of TCGA on lung cancer research under 4 themes: diagnostic markers, disease progression markers, novel therapeutic targets, and novel tools.

LUNG CANCERS

Lung cancer is the leading cause of cancer-related death worldwide, accounting for 1.59 million deaths of the 8.2 million total cancer deaths in 2012.[6] There are 2 main types of lung cancer: small cell lung carcinoma and non-small cell lung carcinoma (NSCLC). NSCLC accounts for 85%–90% of lung cancer cases and its 2 largest subtypes are lung squamous cell carcinoma (LUSC) and lung adenocarcinoma (LUAD).[7] LUSC accounts for 25%–30% of all total lung cancer cases, whereas LUAD accounts for 40% of all total lung cancer cases. LUSC tends to be found in the middle of the lungs and is associated with smoking. In contrast, LUAD forms on the periphery of the lungs and may be associated with smoking, but is the most common lung cancer type among nonsmokers.[7]

The lack of effective therapeutic options to treat NSCLC patients contributes to poor outcomes and to the high number of lung cancer deaths. NSCLCs have historically been resistant to traditional platinum-based chemotherapy with response rates of only 20%–50%, compared with advanced small cell lung carcinomas that have chemotherapy response rates of 60%–80%.[8] This resistance has proven to be a challenge as the dose-response curve for advanced NSCLC plateaus at higher doses, resulting in poor prognosis and low 5-year survival rates.[9,10] The increased focus on driver mutations in tumorigenesis provided critical insight for personalized therapeutics in the treatment of NSCLCs. Research into the mutations of the epidermal growth factor receptor (*EGFR*) and the translocation of anaplastic lymphoma kinase (*ALK*) has been fertile ground for the treatment of NSCLCs.[11-13] Targeted therapies for the active oncogenic *EGFR* mutation have significantly improved treatment outcomes, especially for LUAD because this NSCLC type is the most likely to contain *EGFR* aberrations.[11,12,14]

Despite the effective use of *EGFR* tyrosine kinase inhibitor (*EGFR*-TKI) drugs such as gefitinib to treat *EGFR*-mutated cancers, drug resistances from mutations such as the threonine-790 to methionine (T790M) point mutation are major obstacles in the treatment of NSCLCs.[15] Pan-cancer studies have also found LUAD to be highly heterogeneous, adding to the challenge of providing effective treatment.[16] Furthermore, therapies for LUAD have been proven largely ineffective for treating LUSC.[17] A more comprehensive understanding of the genetic pathways of LUAD and LUSC is needed to develop better diagnostic tools and treatment options. The lung cancer project of TCGA has provided the opportunity to study NSCLCs in depth by amassing large amounts of genomic information from hundreds of cases.

TCGA COVERAGE FOR LUNG CANCERS

LUAD and LUSC were 2 of the 3 lung tumors to be studied by TCGA, with the third being mesothelioma. TCGA had 3 main goals for its lung cancer research: (1) to identify the specific genes that differentiate LUSC and LUAD tumors into molecular subgroups, (2) to differentiate broader gene expression patterns between LUAD and LUSC, and (3) to distinguish genomic changes between smokers and nonsmokers. As of June 2015, 521 samples of LUAD and 504 samples of LUSC