Available Resources and Challenges for the Clinical Annotation of Somatic Variations

Catherine I. Dumur, PhD

Next-generation sequencing (NGS) has become an important tool for identifying clinically relevant variants in both inherited disorders and oncology. Variants annotation that enables the creation of meaningful clinical reports often requires mining multiple publicly available databases. There are a number of such resources that have been designed to catalog and mine a plethora of germline variants or mutations. However, when analyzing tumor specimens in clinical settings, one may need to use different or ancillary resources that are specific for somatic variants or actionable mutations that may have clinical or treatment implications. The purpose of this review is to recapitulate the state of the art of somatic variation databases, which can aid in the clinical interpretation of NGS-based assays in oncology. In addition, the current need for collating various annotation sources into one-stop solutions to facilitate faster query execution and better integration into existing laboratory information systems are discussed. ***Cancer (Cancer Cytopathol)* 2014;122:730-6.** © 2014 The Authors. *Cancer Cytopathology* published by Wiley Periodicals, Inc. on behalf of *American Cancer Society*. This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

KEY WORDS: public databases; somatic variants; actionable mutations; clinical genomic reports; cancer targeted therapies.

INTRODUCTION

The recent technological advances in next-generation sequencing (NGS) and its applications in the field of oncology have allowed the discovery of variants that may belong to the category of somatic mutations. Such findings may enhance the development of targeted cancer therapeutics, which could benefit individuals with tumors harboring such mutations.[1] Acquired mutations in genes encoding for proteins involved in cell growth, proliferation, and survival signaling pathways can be "drivers," that is, can cause cancer or disease or can be related to disease, or "passenger," mutations. Thus, molecular testing to identify somatic mutations in clinical specimens to assess patient eligibility for targeted therapies has become increasingly important in the management of oncology patients.[2]

The advent of benchtop sequencers has allowed the rapid application of molecular testing for somatic mutations in clinical settings.[3] Thus, amplifying discrete or targeted regions of the genome has allowed for the development of panels of "amplicon sequencing." As an example, the Ion AmpliSeq Cancer Hotspot Panel v2 (Life Technologies, Carlsbad, CA), which targets 207 exonic regions across 50 cancer-relevant genes, is producing robust results starting from 1 to 10 ng of DNA isolated from formalin-fixed, paraffin embedded (FFPE) specimens. Such an assay can yield up to 1 Gb of DNA sequences, depending on the chip used to run the sequencing reaction, in short DNA fragments. Similarly, the TruSeq Amplicon — Cancer Panel (Illumina,

Corresponding author: Catherine I. Dumur, PhD, Department of Pathology, Virginia Commonwealth University, Clinical Support Center, Room 247, 403 North 13th Street, Richmond, VA 23298; Fax: (804) 827-4738; cdumur@mcvh-vcu.edu

Department of Pathology, Virginia Commonwealth University, Richmond, Virginia

Received: June 11, 2014; **Revised:** July 21, 2014; **Accepted:** July 22, 2014

Published online August 8, 2014 in Wiley Online Library (wileyonlinelibrary.com)

DOI: 10.1002/cncy.21471, wileyonlinelibrary.com

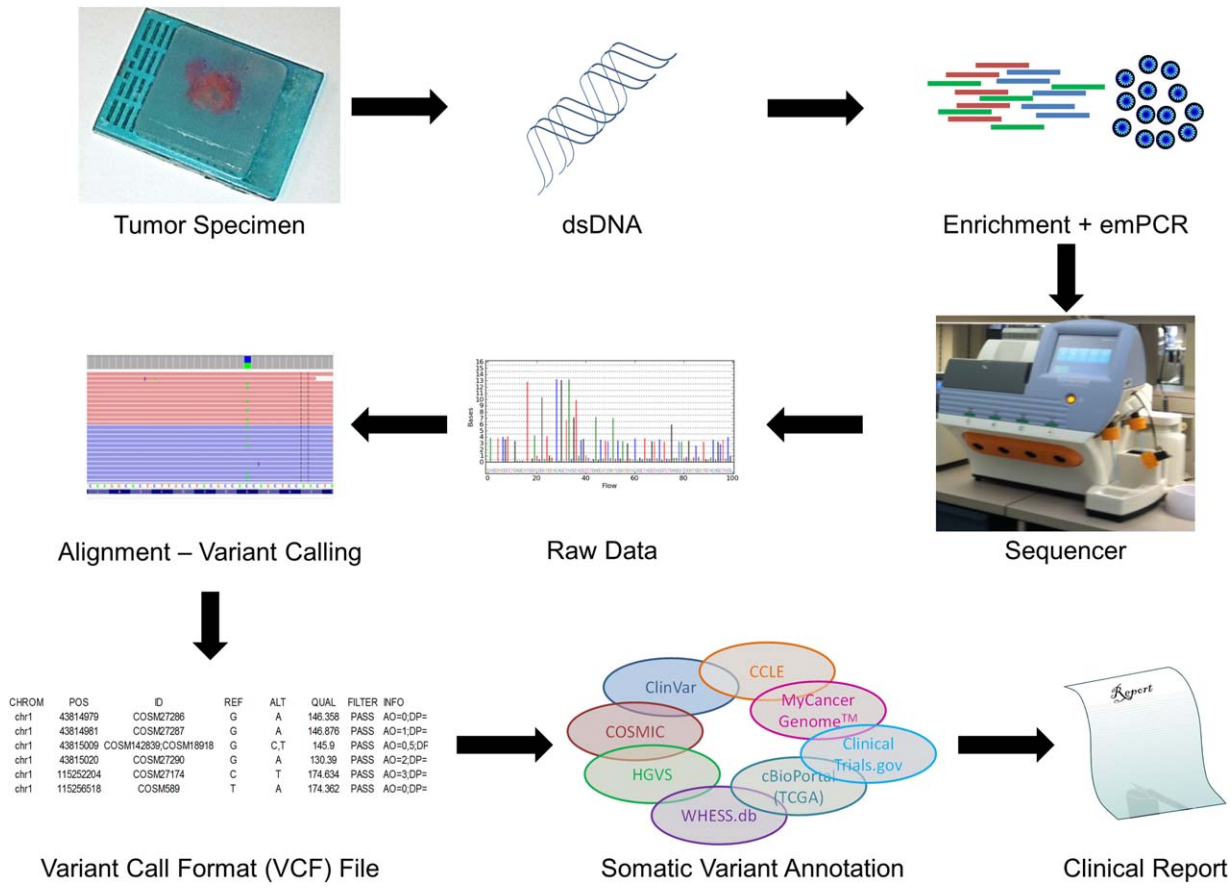

Figure 1. Schematic representation of targeted next-generation sequencing and annotation for tumor specimens. Doubl-stranded DNA (dsDNA) is extracted from FFPE specimens containing tumor cells. Targeted regions are enriched by PCR and clonally amplified by emulsion PCR (emPCR). Such DNA fragments are sequenced by an NGS instrument and aligned to a reference genome. Variations from the reference are called by variant caller algorithms generating a VCF file. Rich variant annotation for somatic mutations can be achieved by querying multiple publicly available databases to create a meaningful clinical report for the management of oncology patients.

Inc., San Diego, CA) assay allows the sequencing of mutational hotspots located in 212 exonic regions corresponding to 48 cancer-related genes from 250 ng of DNA sample.

Several algorithms are then applied to the raw data to align these short reads to a reference genome, assign read and mapping quality scores, and assess those loci that differ from the reference, called variants.[4] These algorithms generate a variant call format (VCF) file,[5] which is a generic format for storing DNA variant data such as single-nucleotide polymorphisms (SNPs), multiple nucleotide polymorphisms (MNPs), insertions (INS), and deletions (DEL), together with quality annotations. VCF files contain variants from a range of positions of the reference genome and are usually stored in a compressed manner. A typical VCF file does not contain information in a way that would be useful for a physician or researcher, such as the transcript and/or gene that contains the variant; the effect, if any, on the encoded protein, such as synonymous, missense, or nonsense mutations; the likelihood that the variant is pathogenic; and the effect on response to targeted therapies. As opposed to human whole-exome sequencing (WES) or whole-genome sequencing (WGS), which can yield nearly 100,000 or 3,600,000 variants[6] per sample, respectively; targeted sequencing for cancer-related gene panels typically yield $<<20$ variants per tumor sample. Even though the medical genomicist processing such VCF files does not have to filter thousands of variants down to a manageable subset, he or she has the important task of distinguishing medically important or actionable variants from the others and reporting them to the treating physician in a meaningful manner (Fig. 1). Ideally, this filtering would be a simple operation of

TABLE 1. Characteristics of Databases for the Annotation of Somatic Variants

	Web URL	HTML Search Links	Login Required	Model-Based Information
ClinVar	http://www.ncbi.nlm.nih.gov/clinvar	Yes	No	No
COSMIC	http://www.sanger.ac.uk/cosmic	Yes	No	No
HGVS	http://www.hgvs.org/mutnomen/recs-DNA.html	No	No	No
WHESS.db	http://genetics.bwh.harvard.edu/pph2/dbsearch.shtml	No	No	Yes (PolyPhen-2)
cBioPortal	http://www.cbioportal.org/public-portal	Yes	No	No
CCLE	http://www.broadinstitute.org/ccle	No	Yes (Free)	No
My Cancer Genome	http://www.mycancergenome.org	No	No	No
ClinicalTrials.gov	http://www.clinicaltrials.gov	Yes	No	No

This review focuses on the publicly available resources for the annotation of somatic variants found by next-generation sequencing (NGS) technologies applied to mutational analyses of tumor specimens in clinical settings. The characteristics of the existing databases as well as the ideal tools needed to comprehensively annotate mutations for the creation of clinically relevant reports for oncology are discussed.

intersecting a VCF file with a comprehensive reference database of medically annotated variants. However, such a resource does not yet exist in a publicly available format, and the medical genomicist has to manually mine a plethora of publicly available and expert curated databases focused on human variant information, such as the Human Gene Mutation Database,[7] the National Center for Biotechnology Information Short Genetic Variations database,[8] and the Online Mendelian Inheritance of Man (OMIM; http://omim.org/), among others.[9]

To date, the majority of such databases are rich in information pertaining to human germline variants, which is very helpful when analyzing WES or WGS data for inherited disorders. However, the analysis and annotation of somatic variants requires the mining of additional databases that help in the interpretation of sequencing data from tumor specimens and allow the creation of clinically relevant reports for clinicians to fine-tune their therapeutic approach in the treatment of oncology patients. The content and capabilities of some of the most clinically relevant databases (Table 1) are reviewed and discussed here.

ClinVar

ClinVar (http://www.ncbi.nlm.nih.gov/clinvar)[10] was launched in 2012, with the first public release in April 2013, and is a freely accessible public archive for reports of the relationship of genomic variations to phenotypes, with the proper supporting evidence. The supporting evidence can arise by 3 methods: clinical testing, research, or literature only (extracted from the literature without modification of authors' statements). The data set in ClinVar includes variants from OMIM, GeneReviews, and some locus-specific databases, as well as clinical and research laboratories. ClinVar can be queried with the genomic coordinates of the allele nucleotide in question. It is important to note that ClinVar does not include uncurated sets of data from genome-wide association studies. Even though ClinVar serves as a central repository for predictions of causation with the 5 standard categories, ranging from benign to pathogenic, widely used for germline variant analysis, it is still a valuable tool for somatic mutation analyses because such mutations are also included in the database. In addition, it is not unusual to find germline variants in tumor samples, and those can be easily assessed in ClinVar. The supporting evidence for the clinical relevance of the variant entries relies on the submitter and can be reviewed by other submitters. Moreover, ClinVar, provides gene information from external sources and links to other relevant databases. In short, ClinVar is a compendium of the current understanding of the relationship between genotypes and medically important phenotypes. This database is not specific for somatic mutations, as it is mostly designed for germline mutations to support the establishment of the clinical validity of human variation.

COSMIC

The catalogue of somatic mutations in cancer, or COSMIC (http://www.sanger.ac.uk/cosmic)[11] was launched in February 2004 by the Cancer Genome Project at the Wellcome Trust Sanger Institute as a novel tool that provides integrated genetic data from cancer genes. Some of the key features of COSMIC are that this database contains information on publications related to mutations found in a variety of samples. Samples include benign neoplasms and other benign proliferations, in situ and invasive tumors, recurrences, metastases, and cancer cell lines. In addition, this database includes samples that have

been found to be negative for mutations during screening to allow for frequency data calculations for mutations in different genes in different cancer types. Mutation information is extracted from the original literature and entered into the COSMIC database. In addition, a histology and tissue ontology has been created for all mutations to be mapped to a single version of each gene, so the data can be queried by tissue type, histology, gene symbol, or variant and can be displayed as a graph or as a table or exported in various formats. Even more importantly, the guidelines to submit to the COSMIC database emphasize the need to use standardized mutation syntax based on the Human Genome Variation Society (HGVS) recommended nomenclature to ensure the usefulness of the information stored in this database. Unlike ClinVar, this database contains mostly information on somatic mutations. As of June 2, 2014, the 69th release of COSMIC (COSMIC v69) contains mutations identified in 27,829 different genes from 999,872 samples and mutant details across a total of 9424 cancer genomes from the international community, including the International Cancer Genome Consortium.

HGVS

The HGVS (http://www.hgvs.org)[12] aims to foster the discovery and characterization of genomic variations, including population distribution and phenotypic associations, by promoting the collection, documentation, and free distribution of genomic variation information and associated clinical annotation. Perhaps the most relevant achievements of the HGVS are the guidelines and recommendations for nomenclature of gene variations that they have issued over the years, starting with 2 articles published in 1993.[13,14] The original suggestions from these publications have been discussed at length, updated multiple times, extended, and ultimately resulted in a set of nomenclature recommendations that have been largely accepted and are applied worldwide in both research and clinical settings. These recommendations are particularly useful when one founds novel variants that have never been described in any of the above-mentioned databases. When that happens, following standardized nomenclature helps in the reporting process of the new variant. In addition, one might want to further analyze the possible impact on protein function of the newly found variant, which can be initially done by using an *in silico* bioinformatics approach such as the SIFT or PolyPhen tools. These tools have been shown to be significantly better at predicting loss-of-function mutations than gain-of-function mutations.[15] Thus, these programs are very useful in prioritizing newly described variants that are likely to cause a loss of protein function. In short, abiding by the HGVS guidelines and recommendations facilitates the standardization of reporting novel somatic variants.

PolyPhen-2 and the WHESS Database

PolyPhen-2 (Polymorphism Phenotyping v2) is a newer iteration of the previous PolyPhen tool, which predicts the possible impact of an amino acid substitution on the structure and function of a human protein using physical and comparative considerations.[16] This tool can be accessed by visiting the following website: http://genetics.bwh.harvard.edu/pph2. There, in addition to using the bioinformatics tool, one can access the WHESS.db, a database of a precomputed set of PolyPhen-2 predictions for whole human exome sequence space (WHESS). This database contains inferred functional annotations for ~150,000,000 single-nucleotide nonsynonymous (missense) codon changes enumerated for each CDS codon position in the exons of 43,043 UCSC *knownGene* transcripts, based on the February 2009 assembly of the human genome, the Genome Reference Consortium Human Reference 37 (GRCh37/hg19) with maximum sequence overlap and identity to known UniProtKB proteins. The WHESS.db can be queried using the genomic coordinates for the allele nucleotide in question and is an excellent tool to identify putative somatic or likely pathogenic variants, precalculated with PolyPhen-2, when analyzing newly found variants in tumor specimens.

cBioPortal and TCGA

The cBioPortal for Cancer Genomics (http://www.cbioportal.org/public-portal/)[17,18] provides access to visualization, analysis, and download of large-scale cancer genomics data sets. Such data sets have been generated by The Cancer Genome Atlas (TCGA) program, which began as a 3-year pilot project in 2006 funded by the National Cancer Institute (NCI) and the National Human Genome Research Institute. As a result of the TCGA pilot project, it was established that an atlas of changes could be created for specific cancer types from different organs/systems, including breast, central nervous system, and endocrine, among others. As of May 21, 2014, the cBioPortal contains data for 17,584 tumor

samples from 69 different cancer genomics studies. This program also allowed the development of an infrastructure for making the data publicly available to researchers and clinical genomicists. Thus, one can mine the TCGA data sets using HGVS mutation nomenclature, or the gene symbol, on the cBioPortal for Cancer Genomics and identify somatic variants that may not as yet been entered into the COSMIC database. One of the striking features of the cBioPortal is that one can access comprehensive, integrated analyses, including but not limited to copy number variation, mRNA and miRNA expression, promoter methylation, and DNA sequence/mutation analysis from only the 69 cancer data sets available from the TCGA project.

CCLE

The Cancer Cell Line Encyclopedia (CCLE) project (http://www.broadinstitute.org/ccle)[19] is a collaboration between the Broad Institute and the Novartis Institutes for Biomedical Research and its Genomics Institute of the Novartis Research Foundation to conduct a detailed genetic and pharmacologic characterization of a large panel of human cancer cell lines. The CCLE provides public access to genomic data, analysis, and visualization for more than 1000 cell lines. To access mutation data related to genes in the database, one must become a registered user for free to log into the CCLE project. To date, the CCLE has 17,683 registered users, and contains genomic data from 1074 samples. The CCLE database can be easily queried by cell line, gene symbol, or tissue type. This database is very useful for the analysis of novel variants identified in tumor specimens in a clinical setting. Even though the information in this database is related to human cell lines, most of these cell lines arose from patient samples at some point, and they may help to shed some light on the nature of the novel variant, along with other supporting evidence from other databases and/or publications.

My Cancer Genome

My Cancer Genome (http://www.mycancergenome.org)[20] is a knowledge database that provides cancer-specific genetic-related information for the era of personalized cancer medicine. Launched in January 2011 by the Vanderbilt-Ingram Cancer Center as a web-based precision cancer medicine knowledge database, this resource gives physicians and physician-scientists access to genetic and treatment information on well-characterized somatic mutations on cancer-related genes. Searches can be conducted by gene, disease (cancer type) or variant (eg, somatic mutation). This is an excellent resource to obtain weekly updated information on the rapidly expanding list of somatic variants that impact different cancers, as well as current information on various mutation- or gene-specific therapies, including those in clinical trials, either locally, nationally, or internationally. The information entered in this database comes from a variety of contributors form the United States, Europe, and Australia. Contributors are clinical and/or scientific experts in their specific disease area. They contribute content based on their areas of knowledge and expertise. In addition, My Cancer Genome provides information from external sources and links to other relevant databases, such as COSMIC, NCI, cBioPortal, and ClinicalTrials.gov, among others. The My Cancer Genome is an excellent one-stop resource for well-known and characterized somatic variants that can be used to generate meaningful reports for treating oncologists. My Cancer Genome contains excellent genotype:-phenotype correlation data for well-known somatic mutations, similar to how ClinVar relates to germline mutations.

ClinicalTrials.gov

ClinicalTrials.gov (http://www.clinicaltrials.gov)[21] is a "registry" and "results database" of publicly and privately supported clinical studies of human participants conducted around the world. This database was launched in February 2000 and currently lists 168,454 studies from all 50 states in the United States and 187 other countries. ClinicalTrials.gov was created as a result of the Food and Drug Administration Modernization Act of 1997, which required the US Department of Health and Human Services, through the National Institues of Health (NIH), to establish a registry of clinical trials information. The NIH and the Food and Drug Administration jointly developed the website. This web site is maintained by the National Library of Medicine at the NIH and is updated daily. The information on ClinicalTrials.gov is provided and updated by the sponsor or principal investigator of the clinical study. Clinical trials are generally registered on the website when they start, and the information is updated throughout the study. In some cases, results of the study are submitted after the clinical trial ends. The website can

be queried by gene symbol, variant, or disease (cancer type).

Future Resources Needed

Currently, there is a growing need for a one-stop, comprehensive reference database of clinically relevant variants that can be easily cross-referenced to NGS-generated data, such as VCF files. A resource with links to other databases containing additional information would be desirable, but querying multiple sources for each variant should be kept to a minimum. The creation of such a resource that will serve as a comprehensive reference database will require the constant submission of all available data for each variant, as well as new data and annotations over time.

Currently, some clinical laboratories have chosen to create local databases that store rich annotation information on a set of sequence variants that have been encountered by the laboratory at least once during mutational analysis testing. These in-house variant knowledge databases often directly import annotations from the various publically available variant databases described above. Collating various annotation sources into smaller subsets of variants into these local variant knowledge databases facilitates faster query execution and allows for accurate and rapid creation of meaningful clinical reports as well as training such knowledge databases with annotation on newly found variants for subsequent reports.[22,23] For somatic mutation analyses, such local databases can be implemented and require the joint expertise of bioinformaticians, programmers, molecular pathologists, and oncologists. In addition, there is currently a gap between NGS results output and the generation of a comprehensive report in the existing laboratory information systems. The currently laborious nature of variant annotation has further intensified the urgent need to close that gap to minimize errors and to maximize efficiency.

Conclusions

The advent of next-generation sequencing (NGS) technologies has revolutionized the detection and reporting of somatic mutations. Currently, clinical and genomic annotation for reporting known or novel somatic variants relies on the mining of multiple publicly available databases. Ideally, a one-stop public, periodically updated, and curated database would be desirable, but such a resource is not currently available. As an alternative, certain laboratories have developed local databases by joining the expertise of bioinformaticians, programmers, molecular pathologists, and oncologists. The demands for clinical data interpretation in clinical settings are urgent. Standardization of protocols that guide how NGS data for somatic variants will be reported and conveyed to the clinician is needed to better utilize this information in the management of oncology patients.

FUNDING SUPPORT

No specific funding was disclosed.

CONFLICT OF INTEREST DISCLOSURES

The author made no disclosures.

REFERENCES

1. Zhao Y, Adjei AA. Targeting oncogenic drivers. *Prog Tumor Res*. 2014;41:1-14.
2. Dumur CI, Idowu MO, Powers CN. Targeting tyrosine kinases in cancer: the converging roles of cytopathology and molecular pathology in the era of genomic medicine. *Cancer Cytopathol*. 2013;121:61-71.
3. Tsongalis GJ, Peterson JD, de Abreu FB, et al. Routine use of the Ion Torrent AmpliSeq Cancer Hotspot Panel for identification of clinically actionable somatic mutations. *Clin Chem Lab Med*. 2014; 52:707-714.
4. Pabinger S, Dander A, Fischer M, et al. A survey of tools for variant analysis of next-generation genome sequencing data. *Brief Bioinform*. 2014;15:256-278.
5. Danecek P, Auton A, Abecasis G, et al. The variant call format and VCFtools. *Bioinformatics*. 2011;27:2156-2158.
6. Abecasis GR, Auton A, Brooks LD, et al. An integrated map of genetic variation from 1,092 human genomes. *Nature*. 2012;491:56-65.
7. Stenson PD, Mort M, Ball EV, et al. The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. *Hum Genet*. 2014;133:1-9.
8. Sherry ST, Ward MH, Kholodov M, et al. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res*. 2001;29:308-311.
9. Johnston JJ, Biesecker LG. Databases of genomic variation and phenotypes: existing resources and future needs. *Hum Mol Genet*. 2013;22:R27-R31.
10. Landrum MJ, Lee JM, Riley GR, et al. ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res*. 2014;42:D980-D985.
11. Bamford S, Dawson E, Forbes S, et al. The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website. *Br J Cancer*. 2004;91:355-358.
12. den Dunnen JT, Antonarakis SE. Mutation nomenclature extensions and suggestions to describe complex mutations: a discussion. *Hum Mutat*. 2000;15:7-12.
13. Beaudet AL, Tsui LC. A suggested nomenclature for designating mutations. *Hum Mutat*. 1993;2:245-248.
14. Beutler E. The designation of mutations. *Am J Hum Genet*. 1993; 53:783-785.
15. Flanagan SE, Patch AM, Ellard S. Using SIFT and PolyPhen to predict loss-of-function and gain-of-function mutations. *Genet Test Mol Biomarkers*. 2010;14:533-537.

16. Adzhubei IA, Schmidt S, Peshkin L, et al. A method and server for predicting damaging missense mutations. *Nat Methods*. 2010;7:248-249.
17. Cerami E, Gao J, Dogrusoz U, et al. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov*. 2012;2:401-404.
18. Gao J, Aksoy BA, Dogrusoz U, et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci Signal*. 2013;6:l1.
19. Barretina J, Caponigro G, Stransky N, et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*. 2012;483:603-607.
20. Pao W. New approaches to targeted therapy in lung cancer. *Proc Am Thorac Soc*. 2012;9:72-73.
21. Lacroix EM, Mehnert R. The US National Library of Medicine in the 21st century: expanding collections, nontraditional formats, new audiences. *Health Info Libr J*. 2002;19:126-132.
22. Roy S, Durso MB, Wald A, et al. SeqReporter: automating next-generation sequencing result interpretation and reporting workflow in a clinical laboratory. *J Mol Diagn*. 2014;16:11-22.
23. Sharma MK, Phillips J, Agarwal S, et al. Clinical genomicist workstation. *AMIA Jt Summits Transl Sci Proc*. 2013;2013:156-157.