The NEW ENGLAND JOURNAL of MEDICINE

Perspective
SEPTEMBER 22, 2016

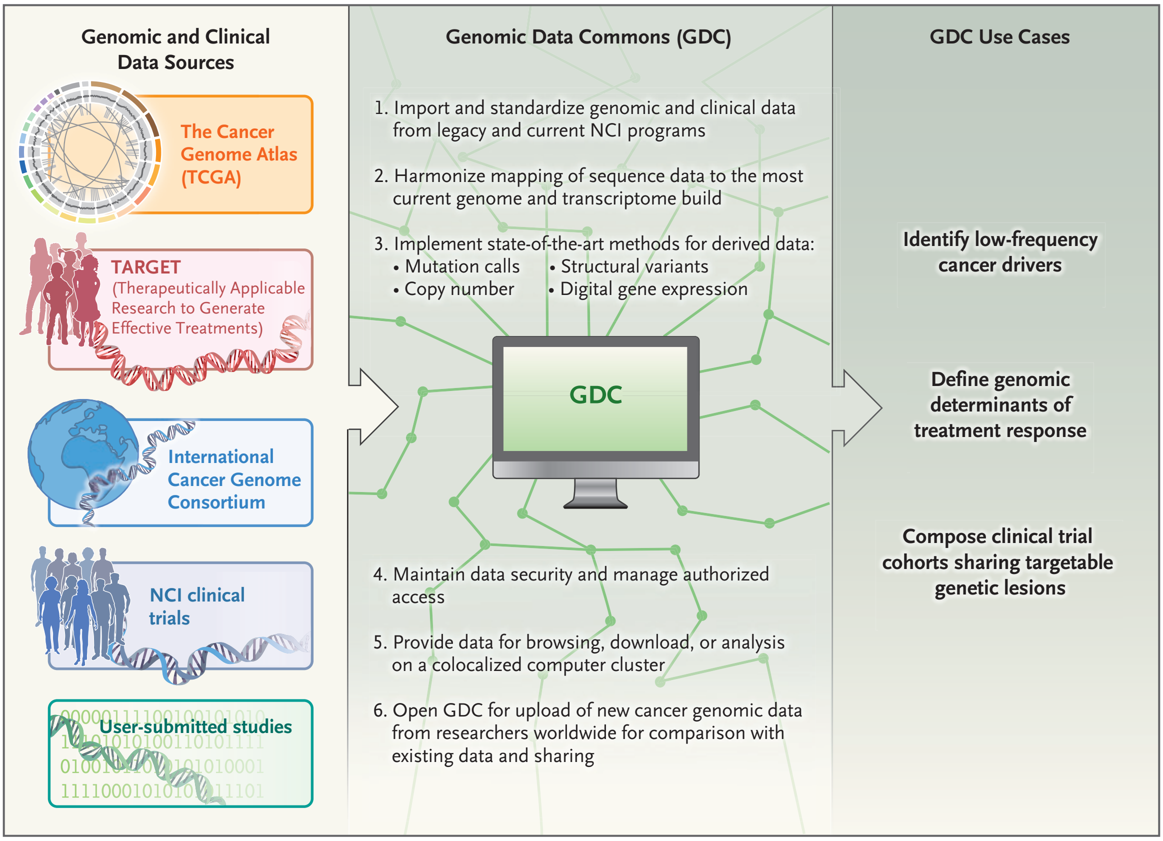
Toward a Shared Vision for Cancer Genomic Data

Robert L. Grossman, Ph.D., Allison P. Heath, Ph.D., Vincent Ferretti, Ph.D., Harold E. Varmus, M.D., Douglas R. Lowy, M.D., Warren A. Kibbe, Ph.D., and Louis M. Staudt, M.D., Ph.D.

For the past 2 years, the National Cancer Institute (NCI), the University of Chicago, the Ontario Institute for Cancer Research, and Leidos Biomedical Research have been developing an information system called the NCI Genomic Data Commons (GDC) (see figure). The GDC will initially contain raw genomic data as well as diagnostic, histologic, and clinical outcome data from NCI-funded projects such as the Cancer Genome Atlas (TCGA) and the Therapeutically Applicable Research to Generate Effective Treatments (TARGET) program. Unlike previous versions of these data sets, the genomic data will be "harmonized" using uniform analytic pipelines to align the raw sequencing data to the genome and identify mutations, copy-number alterations, and gene-expression changes. The research community can access the GDC through an interactive portal (https://gdc-portal.nci.nih.gov), computer systems can interact through the GDC Application Programming Interface, and developers can suggest new features based on GDC open-source code.

An unusual and powerful feature of the GDC is that all researchers will be welcome to submit their cancer genomics data and use the system's computational pipelines, as long as they agree to share their data broadly. The GDC will add value to the researcher's own project by providing access to state-of-the-art bioinformatics tools, and the researcher's data will incrementally increase the interpretive power of the GDC. The system enables researchers to meet the data-access standards for publication in scientific journals and the requirements of the National Institutes of Health (NIH) Genomic Data Sharing policy, and it uses the database of Genotypes and Phenotypes (dbGaP) system to ensure proper data use as specified in informed-consent documents.

Clearly, data sharing will have a pivotal role in precision oncology. A worthy goal set forth by the Institute of Medicine is to develop a new taxonomy of disease based on molecular pathogenesis. By facilitating the sharing of cancer genomic and clinical data, the GDC will gather the information needed by the research and medical community to build a new molecular taxonomy of cancer that has clinical utility. Analysis of current cancer genomic data sets suggests that we are still far from uncovering all the genetic alterations that promote malignant phenotypes, which are known as


Genomic and Clinical Data Sources
The Cancer Genome Atlas (TCGA)
TARGET (Therapeutically Applicable Research to Generate Effective Treatments)
International Cancer Genome Consortium
NCI clinical trials
User-submitted studies
Genomic Data Commons (GDC)
GDC
GDC Use Cases
Identify low-frequency cancer drivers
Define genomic determinants of treatment response
Compose clinical trial cohorts sharing targetable genetic lesions

1. Import and standardize genomic and clinical data from legacy and current NCI programs
2. Harmonize mapping of sequence data to the most current genome and transcriptome build
3. Implement state-of-the-art methods for derived data:
   - Mutation calls
   - Copy number
   - Structural variants
   - Digital gene expression
4. Maintain data security and manage authorized access
5. Provide data for browsing, download, or analysis on a colocalized computer cluster
6. Open GDC for upload of new cancer genomic data from researchers worldwide for comparison with existing data and sharing

Figure 1. Functionality and Utility of the National Cancer Institute Genomic Data Commons (GDC).

The GDC will accept cancer genomic and clinical data from a number of different sources, harmonize the data using consistent bioinformatic pipelines, and allow users to make discoveries regarding the genetic basis of cancer and its impact in the clinic and potentially to identify patients whose tumor profiles make them eligible for particular clinical trials.

cancer "drivers." These calculations suggest that in order to discover genes that acquire driver mutations in 2% or more of patients with cancer, more than 100,000 cancers need to be analyzed.[1] The prevalence of cancer drivers follows a long-tail distribution, meaning that the driver events causing disease in many patients with cancer are not prevalent alleles, such as *BRAF* V600E in melanoma, but are, rather, rare alleles, many of which have not yet been described.

The need to identify such rare drivers of cancer is clear: for any given patient with cancer, detection of a rare genetic driver may be the key to successful therapy. For example, translocation and overexpression of the *ROS1* gene occurs in roughly 1% of lung adenocarcinomas, and small-molecule ROS1 inhibitors can induce complete or partial responses in many affected patients.[2] The co-occurrence of mutations in rare subgroups of cancers can limit the effectiveness of single targeted drugs such as vemurafenib, but in some cases the problem may be overcome by combinations of drugs.[3] The clinical heterogeneity of human cancers is also driven by epigenetic diversity. For example, the response of diffuse large B-cell lymphomas to targeted agents can be predicted by gene-expression profiles.[4] Hence, multiple genomic methods are required to provide a molecular description of cancer that has maximum clinical import.

From the inception of cancer genomics, the value of data sharing has been evident. First, the complexity of genomic data inevitably means that only a fraction of the insights inherent in the data can be reported in any one publication. Second, researchers cannot realistically generate within any one project all the genomic data necessary to draw important conclusions, but they can enrich their study by reusing genomic data from other projects.

However, a major challenge for researchers working with cancer genomic data sets is their sheer size. The TCGA data set alone is over a petabyte in size and consists of more than 575,000 files. Just to download the data using a 10-Gbit-per-second connection would take over 3 weeks. Setting up a secure, compliant infrastructure of sufficient scale to store and analyze the data is not only technically challenging, but also expensive. In 2016, the computing equipment required to analyze the raw TCGA sequencing data costs over $1 million, not including the cost of systems maintenance, security, and compliance that are necessary when working with human genomic and clinical data. The GDC addresses these logistic and economic barriers by democratizing access to cancer genomics data, enabling researchers to bring their hypotheses to the data.

The recent explosion of cancer genome analysis has left in its wake a trail of data ambiguity that must be addressed and rectified. Often, genomic studies are published without the authors' providing raw sequencing data in a public repository, making it impossible to judge the validity of the reported genetic aberrations. Identifying somatic genetic alterations in cancer samples is challenging because of variable contributions of nonmalignant cells, changes in gene copy number, and the presence of tumor subclones. Similarly, the description of copy-number alterations and the quantification of gene expression have not been standardized, posing a problem for researchers trying to compare data from different studies. The GDC addresses these issues by using a harmonization process in which raw sequencing reads are processed through uniform analytic pipelines, and it provides results from multiple analytic methods when there is no single standard. As the human genome sequence is further refined and annotated, and as better analytic pipelines are developed, the GDC will reharmonize its entire genomic content.

The recent explosion of cancer genome analysis has left in its wake a trail of data ambiguity that must be addressed and rectified.

The GDC is the foundation of a multiyear NCI effort to foster a new molecular taxonomy of cancer that provides prognostic information and predicts response or resistance to particular therapies. This project will require the curation of data from clinical trials with embedded genomics and from laboratory experiments that assign biologic phenotypes to particular cancer variants, as well as the development of new methods for integrating multiple clinical and molecular data types. Achievement of this goal will be greatly facilitated by genomic data sharing through the GDC, which will be required for all researchers supported by the NCI and should prove attractive to any investigator whose research would benefit from GDC tools. The genomic and clinical data from NCI-sponsored precision-medicine trials such as the NCI Molecular Analysis for Therapy Choice (MATCH) study will be shared through the GDC, allowing researchers to discern the molecular basis of response and resistance to the many targeted agents being investigated.

The NCI expects investigators to share clinical trial data as outlined in a recent editorial[5] and believes that the GDC will be an appropriate venue for sharing data from NCI-supported clinical trials. As the number of cases in the GDC grows, GDC data could provide evidence of drugs working in cancer subtypes that are too rare to be discerned in smaller clinical trial cohorts.

The GDC could expand rapidly as the acquisition of genomic data becomes routine in the course of cancer care. In time, it may be possible for individual patients to become "cancer information donors" and allow their genomic data to be shared through the GDC. Mechanisms for enabling such donation are being developed under the NIH Precision Medicine Initiative Cohort program. Given appropriate informed-consent systems, the GDC could identify patients with rare molecular subtypes of cancer who could be contacted for potential participation in clinical trials appropriate for their particular cancer.

Clearly, the principles and practice of precision oncology will be accelerated by sharing data from thousands of patients with cancer. We hope that the GDC will be embraced by re-

An audio interview with Jeffrey Drazen is available at NEJM.org

searchers, clinicians, regulatory agencies, patients, and other interested parties as a means to achieve this goal.

Disclosure forms provided by the authors are available at NEJM.org.

From the Center for Data Intensive Science, University of Chicago, Chicago (R.L.G., A.P.H.); the Ontario Institute for Cancer Research, Toronto (V.F.); Weill Cornell Medicine, Cornell University, New York (H.E.V.); and the National Cancer Institute, Bethesda, MD (D.R.L., W.A.K., L.M.S.).

1. Lawrence MS, Stojanov P, Mermel CH, et al. Discovery and saturation analysis of cancer genes across 21 tumour types. Nature 2014;505:495-501.

2. Shaw AT, Ou S-HI, Bang Y-J, et al. Crizotinib in *ROS1*-rearranged non–small-cell lung cancer. N Engl J Med 2014;371:1963-71.

3. Larkin J, Ascierto PA, Dréno B, et al. Combined vemurafenib and cobimetinib in *BRAF*-mutated melanoma. N Engl J Med 2014; 371:1867-76.

4. Wilson WH, Young RM, Schmitz R, et al. Targeting B cell receptor signaling with ibrutinib in diffuse large B cell lymphoma. Nat Med 2015;21:922-6.

5. Taichman DB, Backus J, Baethge C, et al. Sharing clinical trial data — a proposal from the International Committee of Medical Journal Editors. N Engl J Med 2016;374: 384-6.

DOI: 10.1056/NEJMp1607591

Incentives for Clinical Trialists to Share Data

Bernard Lo, M.D., and David L. DeMets, Ph.D.

Sharing of data from clinical trials benefits patients by enabling new discoveries, meta-analyses, and confirmation of published results. As the table shows, the European Medicines Agency (EMA), a number of drug companies, and one other trial funder have already implemented data sharing. A comprehensive Institute of Medicine (IOM) report recommends the sorts of data that should be shared, how long after a trial, and under what conditions.[1] The International Committee of Medical Journal Editors (ICMJE) proposes that the analytic data set supporting a published article be shared no later than 6 months after publication.[2] Others propose longer periods of exclusive data access for trialists.[3] The challenges now are to share data effectively and to minimize disruptions to the clinical trials system, including those affecting trialists who devote years to designing, conducting, and analyzing trials.

Many academic clinical trialists have deep concerns about data sharing.[3] They fear that other investigators will gain unfair rewards from their work and that coinvestigators and mentees will no longer have preferred access to data sets in return for working on the trial. But many trials are never published, and many secondary analyses never get done. Data sharing allows other investigators to carry out these analyses, providing the public with new knowledge gleaned from the contributions of trial participants.

Clinical trialists have practical know-how about a data set that facilitates valid secondary analyses. One of us recently returned to a large data set to carry out a secondary analysis 10 years after trial completion. Although his group had acted as the trial's biostatistics center and documented the data set extensively, including keeping statistical programs, he found he had forgotten important features of the data set, such as the rationale for defining derived variables and censoring rules. Reproducing the published results exactly was challenging because the final data set had been slightly updated from the data set used for publication — some additional events had been discovered during trial closeout. Key staff members were no longer available to provide needed details. Fortunately, with sufficient effort, the new analysis was completed successfully.[4]

Trialists tend to document what they need for their own immediate use and not consider what will be needed for later secondary analyses, perhaps even their own. If a statistical center finds it hard to reanalyze its own data set, it would be even more difficult for secondary users working with an unfamiliar data set. Collaborating with the original trialists would help other investigators derive new knowledge from shared data.

Clinical trialists would view data sharing more favorably if their concerns were addressed. First, funders and sponsors could provide resources for clinical trial data sharing.

Second, clinical trialists could be given incentives to share data. Trialists could receive appropriate acknowledgment and academic rewards when other researchers use "their" shared data to publish