Focus on Computer Resources

Cancer Research

The Digital Slide Archive: A Software Platform for Management, Integration, and Analysis of Histology for Cancer Research

David A. Gutman[1,2,3], Mohammed Khalilia[1], Sanghoon Lee[1], Michael Nalisnik[2], Zach Mullen[4], Jonathan Beezley[4], Deepak R. Chittajallu[4], David Manthey[4], and Lee A.D. Cooper[2,3,5]

Abstract

Tissue-based cancer studies can generate large amounts of histology data in the form of glass slides. These slides contain important diagnostic, prognostic, and biological information and can be digitized into expansive and high-resolution whole-slide images using slide-scanning devices. Effectively utilizing digital pathology data in cancer research requires the ability to manage, visualize, share, and perform quantitative analysis on these large amounts of image data, tasks that are often complex and difficult for investigators with the current state of commercial digital pathology software. In this article, we describe the Digital Slide Archive (DSA), an open-source web-based platform for digital pathology. DSA allows investigators to manage large collections of histologic images and integrate them with clinical and genomic metadata. The open-source model enables DSA to be extended to provide additional capabilities. *Cancer Res; 77(21); e75–78. ©2017 AACR.*

Introduction

Advances in imaging technology have led to an increase in the availability of digital pathology data. Slide-scanning microscopes are capable of digitizing entire histologic sections at 40× objective magnification, generating detailed images that capture tissue microenvironments and cytologic details in high resolution. These images can reveal important information about cancer phenomena like immune response and angiogenesis, can be used to measure expression and localization of proteins, and grade the extent of disease progression. Images produced by slide-scanning microscopes are often several gigabytes in size and can be produced within minutes, resulting in large volumes of imaging data.

There are multiple commercial tools available that enable users to manage, visualize, and analyze collections of digital pathology images. These tools are frequently utilized in pathology cores and shared resources, but have considerable costs and are not customizable or extensible. Commercial tools often require additional fees to increase storage capacity, or to "unlock" additional hardware processors for accelerating image analysis pipelines. Although these tools address many of the frequent needs of cancer researchers, they are closed source and cannot be extended or modified to address specific user needs. There is a need for open-source software in this area that can be developed and maintained by a user community, but the challenges of dealing with digital pathology have made this difficult to accomplish. Digital pathology images often contain a billion+ pixels each, and cancer studies may generate hundreds or thousands of such images. Enabling the management, visualization, and analysis of digital pathology image datasets requires considerable software infrastructure and presents significant challenges for software engineering.

In 2013, we developed the Cancer Digital Slide Archive (CDSA) as an online resource (http://cancer.digitalslidearchive.net) to provide access to images associated with specimens from The Cancer Genome Atlas (TCGA). In addition to the extensive genomic and clinical data collected by TCGA, the Biospecimen Core Resource collected digital pathology images of both frozen and formalin-fixed paraffin-embedded diagnostic sections derived from specimens submitted by Tissue Source Sites. Frozen specimens were collected for quality assurance to assess the tumor purity and extent of necrosis in tissue materials used for genomic analysis. Diagnostic sections were acquired for archival purposes to confirm histologic diagnosis and, in some cases, for evaluation of histologic criteria by TCGA Expert Pathology Committees. Digital pathology images from TCGA have also been used in computational studies that utilize image analysis algorithms to objectively measure histology and to integrate these quantitative measures with genomic and clinical data to improve prognostication and to study molecular pathways associated with histologic phenomena (1–5).

Our motivation behind developing the CDSA was to provide access to these images to the broader research community.

[1]Department of Neurology, Emory University School of Medicine, Atlanta, Georgia. [2]Department of Biomedical Informatics, Emory University School of Medicine, Atlanta, Georgia. [3]Winship Cancer Institute, Emory University, Atlanta, Georgia. [4]Kitware Incorporated, Clifton Park, New York. [5]Department of Biomedical Engineering, Emory University School of Medicine/Georgia Institute of Technology, Atlanta, Georgia.

Note: Supplementary data for this article are available at Cancer Research Online (http://cancerres.aacrjournals.org/).

Corresponding Author: David A. Gutman, Emory University, 201 Dowman Dr, Atlanta, GA 30322. Phone: 404-712-9206; E-mail: dgutman@emory.edu

doi: 10.1158/0008-5472.CAN-17-0629

Although these images are available for download from NCI servers hosting TCGA data, simply viewing them would require downloading terabytes of data and viewing files locally using clients like Aperio Imagescope. The CDSA provides a web-based interface to remotely view and search the TCGA digital pathology images using a web browser, avoiding the need for large downloads and software installation.

Our goal in developing Digital Slide Archive (DSA), which is a more generalized and distributable platform based on the CDSA, was to create a software platform that would enable investigators to create and maintain digital pathology resources for their own tissue-based studies to enable the sharing, annotation, and analysis of their digital pathology data. Enabling users to build and maintain their own web-based archives requires significant software engineering efforts to simplify installation, to document setup and maintenance procedures, and to create the infrastructure needed to manage users, data access permissions, and annotations from human readers and algorithms. We adopted an open-source community development model so that users can customize and extend the DSA project to address their specific research needs.

System Architecture

The DSA is built on a data management toolkit called Girder that is developed and maintained by Kitware (https://data.kitware.com/). Girder provides the infrastructure needed to track images and image metadata, to organize images into collections, to handle user accounts and data access permissions, and to enable data ingestion and administration of DSA installations (Fig. 1A and B). The functionality of Girder can be accessed through its built-in web-based interface or automated programmatically using a RESTful programming interface. The web-based interfaces enable users to easily manage their image collections and metadata, while the RESTful interface enables other applications to automate DSA and integrate digital pathology services into their own applications. Girder also supports limited management of computational analysis workflows and the execution of image analysis pipelines.

The DSA enables users to manage their images in a hierarchical fashion. At the most granular level, there is an image that corresponds to a single file. Multiple images can be assigned to a specimen that may have, for example, a panel of IHC stains, or a sequence of serial sections. A collection can contain multiple images or specimens. Metadata like pathology reports, clinical data, and image acquisition and tissue processing information can be attached to any level of this hierarchy (Fig. 1A). An image can have properties relating to antibodies and staining information or objective magnification, or a pathology report in PDF format can be attached at the specimen level.

A key capability for digital pathology is to support the annotation of image regions by human readers, and the visualization, management, and sharing of these annotations. The DSA enables users to generate annotations within the image viewer using a mouse to create polygons, rectangles, or ellipses. Annotations are organized in layers that correspond, for example, to different tissue compartments, or to the opinions of different readers. Each layer can contain multiple regions, and the layers or regions can be annotated with metadata by users like region name or layer type. Basic annotation information, including timestamps and the reader's user account, are also captured along with the annotation structures and stored within Girder. For backwards compatibility, we support the XML annotation format utilized by Aperio Imagescope, a free Windows-based image viewer with a widespread user base (Fig. 1C).

Whole slide image (WSI) data are represented in pyramid structure containing multiple images at different resolution. For scalability, WSI is partitioned into tiles so it can fit in memory for image analysis and visualization. To handle these operations, a large image plugin has been integrated into Girder, supported by RESTful interface to support WSI operations, such as panning and zooming. The large image plugin relies on the OpenSlide (5) library for handling proprietary digital pathology image file formats. In addition to those core features, large image supports various web-based WSI viewers, including Openseadragon (6), Openlayers (7), Leaflet (8), GeoJS (9), and SlideAtlas (10).

In addition to our work with DSA, we are developing a companion library of image analysis algorithms called HistomicsTK (http://histomicstk.readthedocs.io/en/latest/) to provide image analysis capabilities for DSA installations. HistomicsTK aims to provide the algorithms and pipelines needed for common image analysis tasks like color normalization, basic IHC scoring, cell and region classification, and machine-learning algorithms for prognostic modeling and genomic–histomic integration. In addition to implementations of common algorithms and analysis pipelines, HistomicsTK provides a framework that enables users to integrate their own algorithms and to automatically generate user interface menus for these algorithms within DSA to share them with users.

DSA is an open-source project (https://github.com/DigitalSlideArchive) and is licensed under the Apache 2.0 license to enable the adoption of DSA source by commercial entities. To facilitate easy installation, DSA is available as a prebuilt Docker software container that can be installed on Linux Ubuntu 14.04, 16.04, or Mac OS X systems. The DSA is currently intended for use with deidentified data and images and does not provide the auditing or security capabilities required for HIPAA-compliant storage of private health information. Supplementary Video S1 illustrates the DSA capabilities and installation process.

Discussion

Histology is an important component of tissue-based investigations of cancer, yet utilizing histology imaging in cancer research presents significant challenges. The lack of widespread adoption of a standard image format by most vendors [as described in the Digital Imaging and Communication in Medicine (DICOM) standard Supplement 145 (11)], and the storage requirements present significant challenges to the adoption of this technology. We previously developed the CDSA as a web-based resource that allows users to remotely visualize and search whole-slide histology images associated with TCGA specimens (http://cancer.digitalslidearchive.net/). Users can view more than 10,000 whole-slide images from 32 cancer types, and visualize and browse their clinical data associated with these images, including histologic diagnosis, clinical outcome, treatment information, and deidentified pathology reports in PDF format. An overview of the Digital Slide Archive is displayed in Fig. 1.

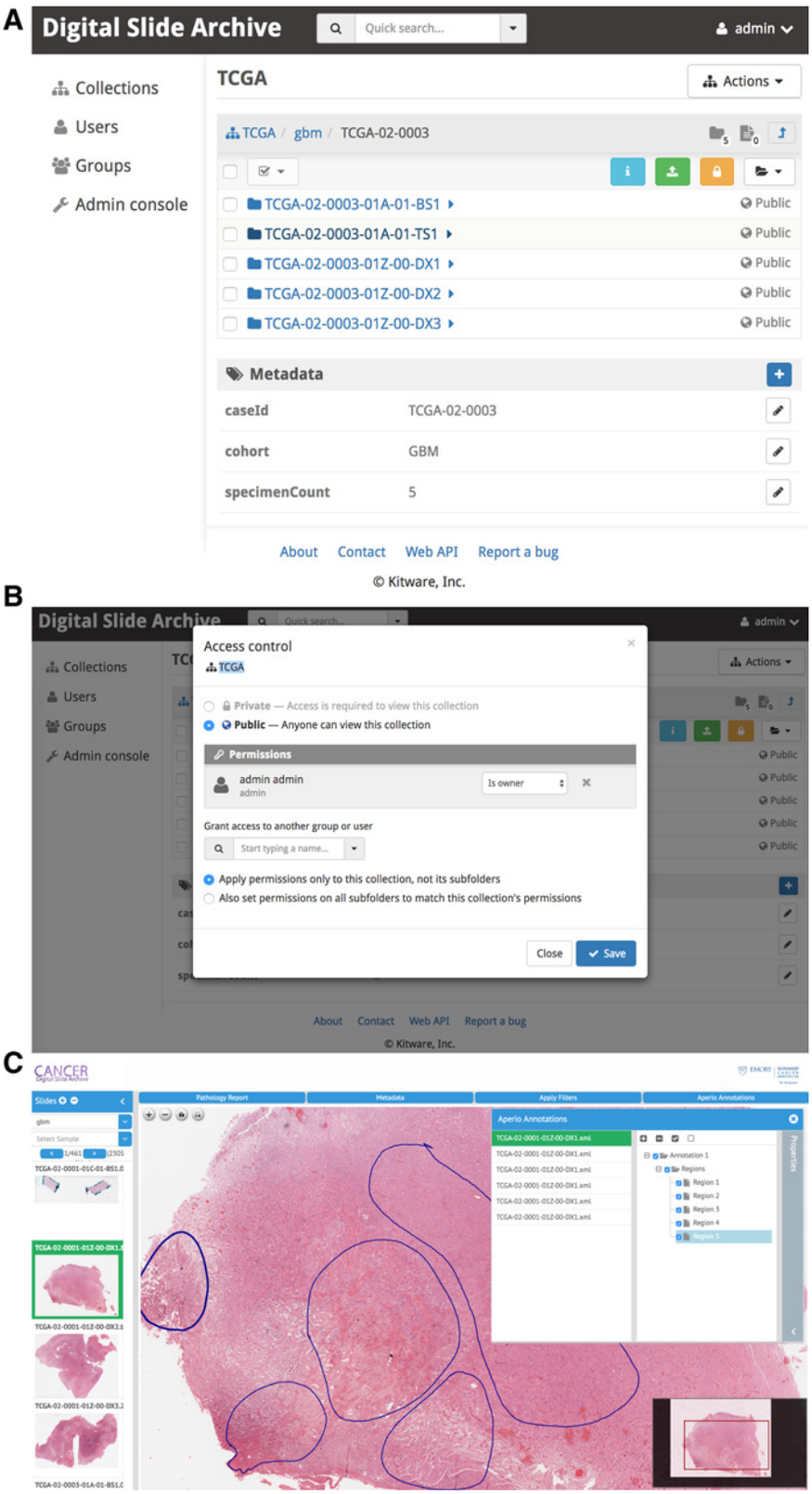

Figure 1.
Overview of the DSA interface. **A,** Girder view showing list of specimens for a given patient (case). **B,** Girder view showing access control widget for a given patient (case). **C,** DSA viewer showing a selected case from GBM and its corresponding Aperio annotation.

The DSA builds upon the CDSA, providing an open-source generalized digital pathology platform so that users can create slide archives to manage, share, and analyze their own histology imaging data. The open-source nature of the DSA and HistomicsTK provides the cancer investigators with a free digital pathology platform, avoiding the need for costly commercial software that is expensive to scale. The DSA and HistomicsTK can be readily extended by others to meet customized user needs, and integrated into other software tools. The DSA was developed in collaboration with Kitware and engineered to make installation, administration, and maintenance of slide archives easy for cancer investigators who may lack technical computing expertise. Support for large-scale analytics and visualization of datasets containing hundreds of millions of objects generated by image analysis is under active development.

Disclosure of Potential Conflicts of Interest

D.A. Gutman has ownership interest (including patents) in and is a consultant/advisory board member for HistoWiz LLC. No potential conflicts of interest were disclosed by the other authors.

Authors' Contributions

Conception and design: D.A. Gutman, M. Khalilia, J. Beezley, L.A.D. Cooper
Development of methodology: D.A. Gutman, M. Khalilia, S. Lee, L.A.D. Cooper
Acquisition of data (provided animals, acquired and managed patients, provided facilities, etc.): D.A. Gutman, M. Khalilia, M. Nalisnik, L.A.D. Cooper
Analysis and interpretation of data (e.g., statistical analysis, biostatistics, computational analysis): D.A. Gutman, M. Nalisnik, D.R. Chittajallu, L.A.D. Cooper
Writing, review, and/or revision of the manuscript: D.A. Gutman, M. Khalilia, D.R. Chittajallu, L.A.D. Cooper
Administrative, technical, or material support (i.e., reporting or organizing data, constructing databases): D.A. Gutman, M. Khalilia, M. Nalisnik, Z. Mullen, J. Beezley, D. Manthey, L.A.D. Cooper
Study supervision: L.A.D. Cooper

Grant Support

This work is supported by the NCI Informatics Technology for Cancer Research grant number U24-CA194362-02.

Received March 9, 2017; revised July 10, 2017; accepted September 25, 2017; published online November 1, 2017.

References

1. Chang H, Han J, Borowsky A, Loss L, Gray JW, Spellman PT, et al. Invariant delineation of nuclear architecture in glioblastoma multiforme for clinical and molecular association. IEEE Trans Med Imaging 2013;32:670–82.
2. Wang C, Pécot T, Zynger DL, Machiraju R, Shapiro CL, Huang K. Identifying survival associated morphological features of triple negative breast cancer using multiple datasets. J Am Med Inform Assoc 2013; 20:680–7.
3. Cooper LAD, Kong J, Gutman DA, Dunn WD, Nalisnik M, Brat DJ. Novel genotype-phenotype associations in human cancers enabled by advanced molecular platforms and computational analysis of whole slide images. Lab Invest 2015;95: 366–76.
4. Kong J, Cooper LAD, Wang F, Gao J, Teodoro G, Scarpace L, et al. Machine-based morphologic analysis of glioblastoma using whole-slide pathology images uncovers clinically relevant molecular correlates. PLoS One 2013;8:e81049.
5. Heng YJ, Lester SC, Tse GM, Factor RE, Allison KH, Collins LC, et al. The molecular basis of breast cancer pathological phenotypes. J Pathol 2017;241:375–91.
6. OpenSeadragon [Internet]. Available from: https://openseadragon.github.io/.
7. OpenLayers. OpenLayers - Documentation [Internet]. [cited 2017 Feb 23]. Available from: http://openlayers.org/en/latest/doc/.
8. Leaflet. Leaflet — an open-source JavaScript library for interactive maps [Internet]. [cited 2017 Feb 23]. Available from: http://leafletjs.com/.
9. OpenGeoscience. OpenGeoscience/geojs: High-performance visualization and interactive data exploration of scientific and geospatial location aware datasets [Internet]. [cited 2017 Feb 23]. Available from: https://github.com/OpenGeoscience/geojs.
10. SlideAtlas. SlideAtlas [Internet]. [cited 2017 Feb 23]. Available from: https://slide-atlas.org/.
11. Singh R, Chubb L, Pantanowitz L, Parwani A. Standardization in digital pathology: supplement 145 of the DICOM standards. J Pathol Inform 2011;2:23.