D3Oncoprint: Stand-Alone Software to Visualize and Dynamically Explore Annotated Genomic Mutation Files

abstract

Purpose Advances in next-generation sequencing technologies have led to a reduction in sequencing costs, which has increased the availability of genomic data sets to many laboratories. Increasing amounts of sequencing data require effective analysis tools to use genomic data for biologic discovery and patient management. Available packages typically require advanced programming knowledge and system administration privileges, or they are Web services that force researchers to work on outside servers.

Methods To support the interactive exploration of genomic data sets on local machines with no programming skills required, we developed D3Oncoprint, a standalone application to visualize and dynamically explore annotated genomic mutation files. D3Oncoprint provides links to curated variants lists from CIViC, My Cancer Genome, OncoKB, and Food and Drug Administration–approved drugs to facilitate the use of genomic data for biomedical discovery and application. D3Oncoprint also includes curated gene lists from BioCarta pathways and FoundationOne cancer panels to explore commonly investigated biologic processes.

Results This software provides a flexible environment to dynamically explore one or more variant mutation profiles provided as input. The focus on interactive visualization with biologic and medical annotation significantly lowers the barriers between complex genomics data and biomedical investigators. We describe how D3Oncoprint helps researchers explore their own data without the need for an extensive computational background.

Conclusion D3Oncoprint is free software for noncommercial use. It is available for download from the Web site of the Biometric Research Program of the Division of Cancer Treatment and Diagnosis at the National Cancer Institute (<https://brb.nci.nih.gov/d3oncoprint>). We believe that this tool provides an important means of empowering researchers to translate information from collected data sets to biologic insights and clinical development.

Clin Cancer Inform. © 2018 by American Society of Clinical Oncology

INTRODUCTION

Continuous advances in next-generation sequencing (NGS) technologies have led to a dramatic reduction in sequencing costs that in turn have increased the availability of genomic data sets to many laboratories. Consequently, increasing amounts of sequencing data require effective analysis tools, particularly those appropriate for the interdisciplinary nature of clinical sequencing projects. Although large institutes typically have a dedicated bioinformatics core unit to support these analyses, investigators from smaller laboratories may struggle to find user-friendly tools for variant analysis.

Available infrastructures and software packages generally require advanced programming

knowledge and system administration privileges, or they may take advantage of Web service architectures that force researchers to work on external servers that may be at risk for potential data exposure. To support the interactive exploration of genomic data sets on local machines and with no programming skills required, we developed D3Oncoprint, a standalone software package for visualization and dynamic exploration of annotated genomic mutation files across groups of samples. We describe how our freely available software can help researchers to explore their own data without the need for an extensive computational background.

The key elements of D3Oncoprint are the following:

Alida Palmisano
Yingdong Zhao
Richard M. Simon

Author affiliations and support information (if applicable) appear at the end of this article.

Corresponding author:

Richard M. Simon, DSc, Biometric Research Program, Division of Cancer Treatment and Diagnosis, National Cancer Institute, 9609 Medical Center Dr, Rockville, MD 20850; e-mail: rsimon@nih.gov.

- The ability to take as input annotated variant files. These can be tab-delimited text files with column headers or variant call format (VCF) file output by popular annotation pipelines.
- The availability of curated variant lists to facilitate the discovery of meaningful biologic information hidden in the genomics data. The curated variants include lists from Clinical Interpretation of Variants in Cancer (CIViC),¹ My Cancer Genome,² OncoKB,³ and Food and Drug Administration (FDA)–approved drugs. D3Oncoprint also allows the use of a customizable user-defined variant list.
- User-controlled selection of the information displayed in the final dynamic plot/table. All selections are performed by using an intuitive graphic interface.
- OncoPrint-style⁴ graphical representation of genetic mutations with additional interactive capabilities. The dynamic exploration features are available through clicks and buttons.
- The availability of curated gene lists from BioCarta pathways and FoundationOne Cancer panels to explore commonly investigated biologic processes.
- A concise textual representation of genetic mutations in a table format with additional interactive capabilities (eg, sorting, filtering, links to external resources such as GeneCards and Catalog of Somatic Mutations in Cancer [COSMIC]).
- The ability to export the D3Oncoprint dynamic view to image files (PNG [portable network graphics], PDF [portable document format], and SVG [scalable vector graphics]) for presentations and publications.
- The ability to export the curated variant table to CSV (comma-separated values) or Excel [Microsoft Corporation, Redmond, WA] files (for additional external analysis).

Related work. The increased availability of tumor sequencing has led to the development of many algorithms for identifying variants from NGS data, predicting which variants have functional effects on protein products and predicting which variants are recurrent across tumors and are likely

to represent driver mutations. Notable examples include VCFtools,⁵ GEMINI,⁶ SNVerGUI,⁷ BrowseVCF,⁸ Ingenuity Pathway Analysis (www.ingenuity.com), VariantStudio (www.illumina.com/products/by-type/informatics-products/variant-studio.html), FireBrowse (<http://firebrowse.org>), and cBioPortal.⁴

Fewer software tools have been developed to support the downstream analysis of annotated variants. Usually large numbers of variants are identified in genomic data sets, and assistance from software tools is necessary to highlight results of clinical interest.

One main disadvantage of some variant analysis tools is that they require advanced bioinformatics programming skills. Others are offered as Web servers, which create issues for researchers who want to analyze their own data in-house. Finally, commercial solutions, although attractive for big centers, may not be a feasible option for small groups.

The main focus of D3Oncoprint is to provide a concise graphical representation of the mutation profiles of groups of samples. Historically, heat maps have been a widely used tool for exploring genomics data; for example, they have been popular in the field of gene expression analysis.⁹ Recently, authors of the cBioPortal project⁴ used heat maps to visually summarize genomic alterations and to reveal correlations in multidimensional genomics data.¹⁰ In this variation of heat maps called OncoPrint, individual genes are represented as rows and individual samples as columns. Colors are used to visually encode genomic alterations, including somatic mutations and copy number changes. This format can be useful for identifying trends such as mutual exclusivity or co-occurrence between gene pairs. These become easily identifiable in the heat maps as clusters of cells equally colored (for cases of mutation co-occurrence) or clusters of cells colored only if other specific cells are not colored (for cases of mutual exclusivity).

The authors of the cBioPortal project provide a Web interface to explore some cancer data sets using this powerful graphical representation. However, it has minimal visual customization options and limited functionality (eg, maximum of 100 genes). Although cBioPortal can be installed on local computers to explore in-house data sets, it requires the assistance of system administrators and a complex infrastructure.

Web exploration of existing cancer data sets can be performed by using the Oncogrid option recently introduced on the Genomic Data Commons (GDC) data portal.¹¹ Oncogrid offers only limited exploration capabilities, and users cannot analyze their own data.

In addition to Web interfaces, R packages such as ComplexHeatmap,¹⁰ GenVisR,¹² and Maftools¹³ allow researchers to create OncoPrint-style plots with their own data. All these packages require advanced programming expertise with R (www.r-project.org) and Bioconductor (www.bioconductor.org) to effectively process commonly used variant data files. In addition, the generated heat map is a static image that is difficult to use for the dynamic exploration of genetic mutation profiles.

D3Oncoprint combines the effective graphical representation of OncoPrint with a carefully designed interface that starts with commonly used variant file formats. D3Oncoprint generates a heat map embedded in a hypertext markup language (HTML) document that can be displayed on modern browsers. Every operation is carried out locally on the user's computer without a requirement for administrative privileges. The heat map can be explored dynamically by taking advantage of the powerful interactive capabilities of the JavaScript library D3 (<https://d3js.org>). The HTML page includes a table with sort-and-filter capabilities to help researchers to also interrogate the genomic information in their input data. Both the heat map and the table can be exported to images and spreadsheets for additional analysis or publication. The intuitive input interface and the interactive output format make complex genomics profiles accessible to researchers and clinicians without requiring bioinformatics expertise, thus facilitating biologic discoveries.

METHODS

D3Oncoprint software is free for noncommercial use. It can be downloaded from the Biometric Research Program (BRP) Web site (<https://brb.nci.nih.gov/d3oncoprint>).

D3Oncoprint consists of an intuitive Java interface that guides the user through five steps to select annotated variant files and to generate the output HTML page (Fig 1). The generated document uses popular JavaScript libraries (eg, D3

and DataTables [<https://datatables.net>]) to present an interactive visual summary of the mutations found in the variant files.

The software is distributed as a Java archive executable by standard Java environment (JRE version 1.8 or above; Oracle, <https://www.oracle.com/java/index.html>). D3Oncoprint runs natively on multiple operating systems without the need for administrative privileges. A detailed user manual that describes D3Oncoprint features is available on the software download Web page.¹⁴

Every operation is carried out locally on the user's computer by using any modern browser that supports JavaScript, SVG, and CSS3 (cascading style sheets) transitions (eg, Google Chrome [Google, Mountain View, CA], Firefox [Mozilla, Mountain View, CA]). When launched, D3Oncoprint connects to the BRP server to look for software or supporting files updates. If new versions are found, the user can update the local files to reflect the most recent versions. We plan to provide maintenance support for the software, including regular updates of curated annotation files. The update of those support files is optional. D3Oncoprint works in environments not connected to the Web because initial copies of all required files are included in the distributed package.

D3Oncoprint User Interface

The D3Oncoprint user interface comprises two tabs. The first tab is used to select the input data (annotated variant files and phenotypes), and the second is used to select which information will be included in the generated HTML page (Fig 2). The user starts by providing a name for the project and selecting the directory where the output HTML file (with the supporting libraries) will be stored. Next, the user selects the annotated variant files from the local machine. Using the "Add input directory button, the researcher can select the directory that contains the desired variant files. Multiple directories can be added in the same way.

D3Oncoprint requires that variants for each sample be stored in a separate file, and file names are used as sample identifiers. All files must be in the same format (tab delimited or VCF format). The variant files can be annotated by various annotation pipelines (eg, ANNOVAR [<http://annovar.openbioinformatics.org/en/latest/>]

Fig 1. In five steps, the D3Oncoprint Java interface collects the required information from the user and generates a local hypertext markup language (HTML) page with the dynamic display of the variant information and sample phenotypes.

or SnpEff [<http://snpeff.sourceforge.net>]]. To process the files, D3Oncoprint only requires that a gene name field, a variant type field, and an amino acid change field are present in the file. The two formats accepted as input are VCF (version 4.0⁵) and tab delimited (refer to the user manual¹⁴ for specifics on the file formats and step-by-step instructions on every part of the graphical interface). If errors or inconsistencies are found, the user is notified with warning messages, which also are recorded in an error log.

The software allows the user to upload a phenotype file, which is a tab-delimited text file where the first column is the sample ID and all other columns (unlimited number) represent different

phenotype information of interest (eg, age, race, histology, stage). These labels are added as colored rows in the D3 heat map output to allow the user to sort the samples according to these phenotypes and to explore information hidden in the genomics data.

After the input information in the first tab is complete (Fig 2A), the user can click the "Load and validate files" button to import the information contained in the headers of the annotated variant files, and all the existing column/fields will be displayed in the second tab (Fig 2B). The user must select which of the available column/fields contains the gene symbol, which contains the variant type, and which contains the amino

Fig 2. (A) Data input tab used to name the project and select the input files from the file system. (B) Data columns from the input files can be selected as the appropriate field to generate the interactive plots. (C) Example of a generated HTML page with an interactive heat map, exploration buttons, and a dynamic data table that displays all variants across the input samples. See Methods and Results for details. MAF, mutation annotation format; TCGA-READ, The Cancer Genome Atlas-Rectum Adenocarcinoma; VCF, variant call format.

acid change code (refer to the user manual¹⁴ for specifics). The correct identification of these fields is critical to build and color the D3 heat map correctly.

Only those three fields are required, but the generated HTML page can include additional information from the variant files. By default, chromosome number, start position, reference, and alternate alleles are added to the list because these fields typically contain valuable information for downstream analysis. The user can decide to add or remove any field of interest. The number of included columns may influence the performance of the HTML page; therefore, we suggest that the user starts with few columns and adds more in the follow-up analysis. After the user clicks the "Create interactive viewer" button (Fig 2B), D3Oncoprint parses all the input variant files, extracts the selected information, and outputs an HTML file (together with all

the required JavaScript libraries) in the specified project directory (Fig 2C).

D3Oncoprint automatically opens the generated HTML page in the default browser. The page has been optimized for the Google Chrome Web browser. If the user's default browser is not a modern browser that supports JavaScript, SVG, and CSS3 transitions, he or she can locate the HTML file in the output directory and manually open it with a modern browser such as Google Chrome or Firefox. In the Results section, we describe the features included in the interactive HTML page.

D3Oncoprint: Curated Support Files

In addition to displaying the information contained in the input variant files, D3Oncoprint includes curated information used to facilitate the discovery of meaningful biologic information from genomics data. The curated information

distributed with the first version of the software includes the following:

- Gene lists grouped according to specific functions (ie, BioCarta pathways¹⁵ and FoundationOne platform genes¹⁶). Gene lists are available on the HTML page, and they are used to display only the mutated genes that belong to a specific group.
- Web links to COSMIC¹⁷ entries for variants that have COSMIC IDs included in the annotation.
- A CIViC-curated variants list that includes variants of interest from CIViC, an open access, community-driven Web resource.¹
- A My Cancer Genome–curated variants list that includes variants in My Cancer Genome, an up-to-date resource on driver mutations with related therapeutic implications, including available clinical trials.²
- An FDA-curated variants list that includes variants of interest for use of FDA-approved drugs with targeted indications.¹⁸
- An OncoKB-curated variant list that includes actionable variants in OncoKB, a precision oncology resource that contains information about the effects and treatment implications of specific cancer gene alterations.³
- A user-defined variant list in which users can include variants of interest for a specific project so that these can be easily marked and highlighted in the maps and tables. The content can be changed by using any text editor (refer to the user manual¹⁴ for details about the file format). Each variant of interest should indicate at least the gene name and amino acid change. Additional information (external links, description, etc) are optional, but if included, they are embedded in the HTML viewer.

If a variant in a sample is matched with any of these curated lists of variants, an icon appears next to the entry, and a link to the source information is provided for additional investigation.

We plan to provide maintenance support for the software, including regular updates of curated annotation files. The frequency of updates depends on the update frequency of the linked resources. Because most of the listed resources do not provide precise timelines, we plan to

revise our information every 6 months. The updated information is distributed seamlessly to any D3Oncoprint user connected to the Web because the software accesses the BRP server every time it is launched. If new versions are found, the user can get the most updated versions with just a click of a button.

RESULTS

To illustrate the interactive capabilities of D3Oncoprint, we provide an example of the use of the publicly available variant files for the colorectal cancer cohort of The Cancer Genome Atlas (TCGA-Rectum Adenocarcinoma [READ]). The files are hosted in the GDC data portal.¹¹ We downloaded the open access mutation annotation format (MAF) files that contain somatic mutations for the TCGA-READ cohort. MAF files that contain any germline mutation annotations are kept in the controlled access portion of the GDC data portal. For demonstration purposes, we use the MAF file from the Mutect pipeline.¹⁹

A single MAF file for the TCGA-READ cohort contains the somatic mutations for tumors of 137 patients. [Figure 2A](#) shows how we gave a name to the project, selected the directory where we extracted the downloaded MAF file, and selected the downloaded JavaScript Object Notation (JSON) file that contains the phenotype information, which is the format used in the GDC portal to encode phenotype information. D3Oncoprint accepts phenotype information either in tab-delimited format or in JSON format (refer to the user manual¹⁴ for details). After the user clicks the "Load and validate files" button, the software parses these input files. Because of the well-defined and controlled structure of MAF files, D3Oncoprint is able to prepopulate the mandatory fields in the "Specify data columns" tab ([Fig 2B](#)). The system is then ready to generate the HTML interactive map with a click of the "create interactive viewer" button ([Fig 2C](#)).

We can quickly observe that the genes with the largest number of mutations across the samples are *APC*, *TP53*, *TTN*, and *KRAS*, which is consistent with the literature.²⁰ We notice that many of the mutations on *TP53*, *KRAS*, *NRAS*, and *PIK3CA* have additional curated annotation because additional information about them is available in My Cancer Genome, CIViC, and OncoKB. This information is accessible by clicking on the icon of interest in the data table

Fig 3. D3Oncoprint heat maps for the TCGA-READ (The Cancer Genome Atlas-Rectum Adenocarcinoma) cohort. Rows represent genes, and columns represent samples. (A) The default view sorts the genes by total mutation count across all the samples, and it sorts the samples by total mutation count across all the genes. Types of mutations are identified by various colors in the gray heat map (point mutations, insertion, deletions, etc). Moving the cursor on a cell with a mutation displays a pop-up with some information about that mutation. Rows can be dragged up or down by clicking and moving the gene name label. Columns can be dragged left or right by clicking and moving the sample name label. Phenotypes are listed in the lower part of the heat map, and the phenotype label acts like a sorting button; a click of any of these buttons moves the columns to separate the samples according to the selected phenotype value: (B) gender and (C) primary diagnosis code. See Results for details about the interpretation of the various maps. FDA, US Food and Drug Administration.

directly below the heat map (Fig 3A). The gender phenotype button allows the user to re-sort the columns into two groups (female, male) as shown in Fig 3B. From the distribution displayed above the heat map, we can see that a difference in terms of total mutation count exists between female and male, and the group of samples with the highest number of mutations comprises all men. With clicking the primary diagnosis code phenotype button, the columns are sorted into multiple groups (c18.9, c19, c20, undefined, and c80.1) and maintain the previous female/male sorting within each group (Fig 3C). From the distribution displayed above the heat map, we can see that the group with the c20 diagnosis (malignant neoplasm of rectum) has the highest number of mutations for both sexes, and all the male patients with the highest number of mutations have c20 as the primary diagnosis. This example shows the kind of observations that are quickly available with D3Oncoprint. Whether

this result is statistically significant will require additional analysis of the data set. Nevertheless, D3Oncoprint is a powerful tool to unravel biologic discoveries by identifying trends such as mutual exclusivity or trends in subgroups of samples or gene panels, hence providing guidance for future investigations.

DISCUSSION

Over the past decade, next-generation sequencing has emerged as an important tool for the diagnosis and treatment of patients with cancer and other diseases. A deep understanding of the relationships of genomic alterations to one another and to phenotypes in groups of patients is at the core of precision medicine approaches to patient care. To support the interactive exploration of genomic data sets on local machines for users with no programming skills, we developed D3Oncoprint, a stand-alone software tool

to visualize and dynamically explore annotated genomic mutation files across samples.

The software provides intuitive dynamic visualization options that can help researchers to translate information encoded in genomics data into biologic insights and potential clinical applications. The current and future development of D3Oncoprint is directed at including additional

visualization options to make the exploration of the data even more dynamic, increase the amount of input information (eg, copy number alterations), and expand the curated information of variants with important clinical implications.

DOI: <https://doi.org/10.1200/CCI.17.00050>

Published online on ascopubs.org/journal/cci on March 7, 2018.

AUTHOR CONTRIBUTIONS

Conception and design: All authors

Collection and assembly of data: All authors

Data analysis and interpretation: All authors

Manuscript writing: All authors

Final approval of manuscript: All authors

Accountable for all aspects of the work: All authors

Alida Palmisano

No relationship to disclose

Yingdong Zhao

No relationship to disclose

Richard M. Simon

Consulting or Advisory Role: AbbVie

AUTHORS' DISCLOSURES OF POTENTIAL CONFLICTS OF INTEREST

The following represents disclosure information provided by authors of this manuscript. All relationships are considered compensated. Relationships are self-held unless noted. I = Immediate Family Member, Inst = My Institution. Relationships may not relate to the subject matter of this manuscript. For more information about ASCO's conflict of interest policy, please refer to www.asco.org/rwc or ascopubs.org/jco/site/ifc.

ACKNOWLEDGMENT

We thank Suleyman Vural, PhD, and Ting Chen, PhD, for providing us with annotated variant files from annotation pipelines with various formats. We also thank Dmitry Sonkin, PhD, Julia Krushkal, PhD, Jianwen Fang, PhD, Ming-Chung Li, PhD, Lori Long, PhD, and all our colleagues in the Biometric Research Program for helpful discussions and suggestions.

Affiliation

All authors: National Cancer Institute, Rockville, MD.

REFERENCES

1. Griffith M, Spies NC, Krysiak K, et al: CIViC is a community knowledgebase for expert crowdsourcing the clinical interpretation of variants in cancer. *Nat Genet* 49:170-174, 2017
2. Swanton C: My Cancer Genome: A unified genomics and clinical trial portal. *Lancet Oncol* 13:668-669, 2012
3. Chakravarty D, Gao J, Phillips SM, et al: OncoKB: A precision oncology knowledge base. *JCO Precis Oncol* 2017:1-6, 2017
4. Cerami E, Gao J, Dogrusoz U, et al: The cBio cancer genomics portal: An open platform for exploring multidimensional cancer genomics data. *Cancer Discov* 2:401-404, 2012
5. Danecek P, Auton A, Abecasis G, et al: The variant call format and VCFtools. *Bioinformatics* 27:2156-2158, 2011
6. Paila U, Chapman BA, Kirchner R, et al: GEMINI: Integrative exploration of genetic variation and genome annotations. *PLOS Comput Biol* 9:e1003153, 2013
7. Wang W, Hu W, Hou F, et al: SNVerGUI: A desktop tool for variant analysis of next-generation sequencing data. *J Med Genet* 49:753-755, 2012
8. Salatino S, Ramraj V: BrowseVCF: A Web-based application and workflow to quickly prioritise disease-causative variants in VCF files. *Brief Bioinform* 18:774-779, 2016
9. Eisen MB, Spellman PT, Brown PO, et al: Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A* 95:14863-14868, 1998

10. Gu Z, Eils R, Schlesner M: Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* 32:2847-2849, 2016
11. National Cancer Institute: Genomic Data Commons Data Portal. <https://gdc-portal.nci.nih.gov>
12. Skidmore ZL, Wagner AH, Lesurf R, et al: GenVisR: Genomic visualizations in R. *Bioinformatics* 32:3012-3014, 2016
13. Mayakonda A, Koeffler HP: Maftools: Efficient analysis, visualization and summarization of MAF files from large-scale cohort based cancer studies. *bioRxiv* 052662, 2016
14. National Cancer Institute: D3Oncoprint: Standalone software to visualize and dynamically explore annotated genomic mutation files: User manual. <https://brb.nci.nih.gov/d3oncoprint/D3Oncoprint-UserManual.pdf>
15. National Cancer Institute: Cancer Genome Anatomy Project. https://cgap.nci.nih.gov/Pathways/BioCarta_Pathways
16. Foundation Medicine: FoundationOne platform. <http://foundationone.com>
17. Forbes SA, Beare D, Boutselakis H, et al: COSMIC: Somatic cancer genetics at high-resolution. *Nucleic Acids Res* 45:D777-D783, 2017
18. US Food and Drug Administration: Resources for information on approved drugs. <http://www.fda.gov/Drugs/InformationOnDrugs/ApprovedDrugs>
19. Cibulskis K, Lawrence MS, Carter SL, et al: Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat Biotechnol* 31:213-219, 2013
20. Cancer Genome Atlas Network: Comprehensive molecular characterization of human colon and rectal cancer. *Nature* 487:330-337, 2012