D3Oncoprint:
Standalone software to visualize and dynamically explore annotated genomic mutation files

User Manual

Alida Palmisano, Yingdong Zhao, Richard Simon

Biometric Research Program,
Division of Cancer Treatment and Diagnosis (DCTD)
National Cancer Institute,
Rockville, MD, USA

Email inquiries to: alida.palmisano@nih.gov, zhaoy@mail.nih.gov, rsimon@nih.gov

Version 1.0

September 2017

Table of Contents

1 General information

This document is the user manual about the D3Oncoprint software.

Some of the text and figures are presented in a manuscript submitted for publication:

Palmisano A., Zhao Y. and Simon R. "D3Oncoprint: a standalone software to visualize and dynamically explore annotated genomic mutation files", JCO Clinical Cancer Informatics (JCO CCI) (2017)

This document contains a general description of the software system, together with step by step instructions on how to use D3Oncoprint on an example of publicly available datasets from TCGA.

1.1 Abstract

Advances in Next-Generation Sequencing technologies have led to a reduction in sequencing costs, increasing the availability of genomic datasets 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, system administration privileges or they are web services that force researchers to work on outside servers.

In order to support the interactive exploration of genomic datasets 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 FDA approved drug 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 biological processes.

This software provides a flexible environment to dynamically explore one or more variant mutation profiles provided as input. The focus on interactive visualization with biological and medical annotation significantly lowers the barriers between complex genomic data and biomedical investigators. In this document, we describe how D3Oncoprint can help researchers explore their own data, without the need of an extensive computational background.

D3Oncoprint is free software for non-commercial use. It is available for download from the website of the Biometric Research Program (BRP) of the Division of Cancer Treatment & Diagnosis at the National Cancer Institute (https://brb.nci.nih.gov/d3oncoprint/). We believe that this tool will provide important means of empowering researchers to translate the information from the collected data sets to biological insights and clinical developments.

2 Introduction

Continuous advances in Next-Generation Sequencing (NGS) technologies have led to a dramatic reduction in sequencing costs which, in turn, has increased the availability of genomic datasets to many laboratories. Consequently, increasing amounts of sequencing data require effective analysis tools, particularly tools appropriate for the inter-disciplinary nature of clinical sequencing projects. While large institutes typically have dedicated bioinformatics core units to support these analyses, investigators from smaller labs may struggle to find user-friendly tools for variant analysis.

Available infrastructures and software packages generally require advanced programming knowledge, system administration privileges or they may take advantage of web services architectures that force researchers to work on external servers that may be at risk of potential data exposure.

In order to support the interactive exploration of genomic datasets 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.

In this document, we present our freely available software and how it can help researchers explore their own data, without the need of an extensive computational background.

The key elements of our D3Oncoprint software are:

1. The ability to take as input annotated variant files. These can be tab delimited text files with column headers or VCF files output of popular annotation pipelines;
2. The availability of curated variants lists that can facilitate the discovery of meaningful biological information hidden in the genomic data. The curated variants include lists from CIViC (Griffith et al. (2016)), My Cancer Genome (Swanton (2012)), OncoKB (Chakravarty et al. (2017)) and FDA approved drugs). D3Oncoprint allows also the use of a customizable user defined variant list;
3. User controlled selection of the information to be displayed in the final dynamical plot/table. All selections are done using an intuitive graphical interface;
4. Oncoprint-style (Cerami et al. (2012)) graphical representation of the genetic mutations, with additional interactive capabilities. The dynamic exploration features are available through simple clicks and buttons;
5. The availability of curated gene lists from BioCarta pathways and FoundationOne Cancer panels to explore the input data for commonly investigated biological processes;
6. A concise textual representation of the genetic mutations in a table format, with additional interactive capabilities (e.g., sorting, filtering and links to external resources like GeneCards and COSMIC);
7. Export of the D3Oncoprint dynamic view to image files (PNG, PDF and SVG) for presentations and publications;
8. Export of the curated variant table to CSV or Excel files (for further external analysis).

The rest of this document describes in details D3Oncoprint's user interface, input format and features of the interactive viewer.

3 How to use D3Oncoprint

D3Oncoprint has been designed and developed by the Biometric Research Program in the Division of Cancer Treatment and Diagnosis of the National Cancer Institute and is available for download from the BRP website (`https://brb.nci.nih.gov/d3oncoprint/`).

IMPORTANT NOTE: Some operating systems may not allow by default the installation of "unofficial apps" downloaded from the internet. This issue is not specific to the D3Oncoprint executable file, but it depends on the OS security requirements. If the D3Oncoprint executable file doesn't seem to run, please inspect the operating system security warnings and allow the D3Oncoprint application to be downloaded and executed in your system. This issue should only occur once, as OS security settings are saved on your machine.

D3Oncoprint consists of an intuitive Graphic User Interface (GUI) that guides the user through a streamlined 5-steps interface to select annotated variant files and it generates an HTML document that uses popular cross-browser Javascript libraries to present an interactive visual summary of the mutations found in the variant files (see Figure 1).

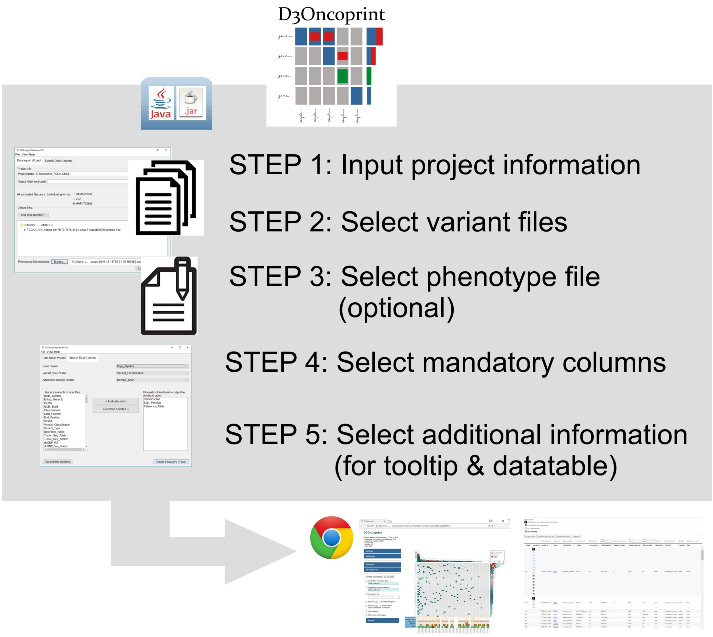

D3Oncoprint
Java
.jar
STEP 1: Input project information
STEP 2: Select variant files
STEP 3: Select phenotype file (optional)
STEP 4: Select mandatory columns
STEP 5: Select additional information (for tooltip & datatable)

Figure 1: *D3Oncoprint Java interface (through five simple steps) collects all the required information from the user and generates a local HTML page with the dynamical display of the input variant information and sample phenotypes. Details on how to use the software are presented in the sections below.*

The GUI is written in Java, version 8, and the main Javascript libraries used are D3 (`https://d3js.org/`) and Datatable (`https://datatables.net/`).

Our software is distributed as a small executable JAR that runs natively on multiple operating systems without the need for administrative privileges and with minimal system requirements (Java(TM) SE Runtime Environment, version 1.8). The distributed package is bundled with some example files and a PDF version of this user manual.

Every operation is carried out locally on the user's computer, using the default browser (we recommend Google Chrome as other browsers do not fully support the Javascript dependencies). When launched, D3Oncoprint tries to connect to BRP server to check if updates of the software or of the supporting files are available: if new versions are found, the user will be able to update the local files to reflect the most recent versions. BRP plans to maintain a curated version of support files, and to provide development support for the software. Updating the support files is completely optional and D3Oncoprint will work in environments not connected to the web as initial copies of all required files are included in the distributed package.

3.1 Starting the graphical user interface

The software is distributed as an executable JAR that can be used without the need of administrative privileges. We recommend to create new empty folder in the file system and save the **D3Oncoprint.jar** in the new directory: this will guarantee that all the support files used by the software (and extracted the first time the jar is executed) will be stored in the same folder and they won't interfere with other user's files.

To launch the GUI the user can simply double click on the D3Oncoprint.jar file and the software window should open (Figure 2). If the Java environment is not setup correctly the double click may not work. In that case, the user need to open a Terminal/Console window, navigate to the directory containing the software jar file and type `java -jar D3Oncoprint.jar`.

Important note: Some operating systems may not allow by default the installation of "unofficial apps" downloaded from the internet. This issue is not specific to the D3Oncoprint executable file, but it depends on the OS security requirements. If the D3Oncoprint executable file doesn't seem to run, please inspect the operating system security warnings and allow the D3Oncoprint application to be downloaded and executed in your system. This issue should only occur once, as OS security settings are saved on your machine.

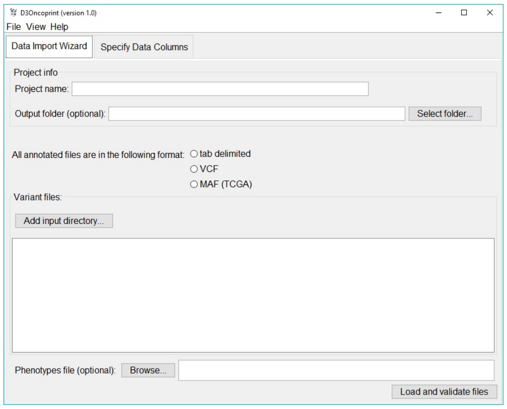


Figure 2: *Main interface of D3Oncoprint. Graphical elements in two tabs guide the user through a sequential process to load the variant files, select the desired output format and create the interactive HTML viewer. Details explained in the text.*

3.2 Menu options

The menu options available in D3Oncoprint are:

1. **File**
   - New: this option clears all text fields, lists and selections in the different tabs of the interface, to allow the user to start with an empty project;
   - Save project: this option asks the user to select a file name to store the current content of the project (all text fields, lists and selections in the different tabs of the interface). The generated file will have '.d3oproject' extension;
   - Load project: this option asks the user to select a file to load a previously saved project. The file must

have the .d3oproject extension and has to be generated by a previous ’Save project’ call;

- Open recent HTML: since users may want to work on different projects, they can generate several interactive viewer HTML pages. With this menu option we give the user the opportunity to directly re-open previously created HTML pages. Note that D3Oncoprint only stores the last 10 projects in this list of recent files, but all the previously generated HTML pages can be directly opened with the browser just by clicking the generated HTML located in the output directory indicated by the user. We provide this option in our menu to facilitate the exploration of multiple recent projects, without the need to remember where they were stored on the local file system;
- Advanced options: this menu opens a popup window where some default options of the software can be easily modified. In this first version of the tool, the two defaults values that can be changed are:
  - (a) the patient ID column that is used to split multi-patient MAF files into single patient files;
  - (b) the maximum number of genes to include in the output HTML page. Since the speed and responsiveness of the generated page depends on browser specific performances, a small number (e.g., 2000) will make sure that the generated page is fast and easier to explore. However if the researcher is interested in interrogating larger gene groups, this number can be increased. Using the -1 value will load all the genes contained in the input annotated variant files.
- Exit: this options quits the program. If there are unsaved changes to the current project, a popup will appear and the user can decide to save (or not) the current project.

2. **View**

- Increase font: this option increases the size of the font for all visual elements of the D3Oncoprint graphical user interface. Note that because of layout constrains, fonts that are too big may cause buttons or textfields to be rearranged in an incorrect fashion (even disappear from the view). Appropriately resize the window if this happens, or decrease the font;
- Decrease font: this option decreases the size of the font for all visual elements of the D3Oncoprint graphical user interface.

3. **Help**

- Updates: every time D3Oncoprint is launched, the software tries to access the BRP server to check for updates. If updated files are found on the server, this menu option will change color (the background will be orange). Clicking this menu option opens a pop-up window listing all the changes found on the server. The user can then decide to update the local files to reflect those changes (by clicking the appropriate button). If the user chooses not to update the files, D3Oncoprint will still be able to generate the interactive viewer HTML, but the information inside these pages could be out-to-date (e.g., old curated variant links);
- Download example demo: this option asks the user to select a directory where download the TCGA-READ data from GDC data portal. After a successful download, the data will be automatically loaded in D3Oncoprint. The user can then just click to the ’Load and validate files’ to parse the TCGA-READ data, and by clicking the ’Create Interactive Viewer’ D3Oncoprint will generate the HTML ouput page with the dynamic heatmap and datatable as displayed in Figure 9 and Figure 10;
- About: a popup with a brief description of the tool is displayed to the user. The pop-up also presents contact information and link to web resources (tutorials, help, etc.).

Note: Most menu options have associated keyboard shortcuts that can trigger the same action. The key combination for each menu is listed next to the menu item and it varies across operating systems.

3.3 Data Import Wizard (Tab 1)

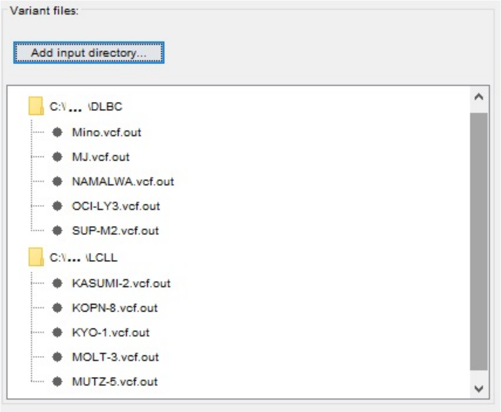
This tab asks the user to setup the main project information.

At the top, the user can specify a name. The name can contain spaces but we suggest simple alphanumeric names because the project name is used as the project folder name and as the name of the HTML output: using symbols and other characters may not work properly on some operating systems. Spaces will be automatically replaced by underscores for file and folder names.

The user can specify a parent output folder, where the interactive HTML viewer will be stored. If left blank, the default location will be the same as D3Oncoprint executable jar. Users need to make sure to have write permission on the selected folder.

The main part of the interface asks the user to provide the input variant files to create the interactive map. All the files must use the same format and three formats are accepted (tab delimited, VCF and MAF). For more details about the difference between those files and how they are processed, please refer to Section 4 on page 17.

After selecting the file format with the radiobuttons, clicking *Add input directory...* will open a file browser that the user can use to navigate the local file system to the directory containing the variant files. All the files contained in that directory will be listed in the hierarchical tree structure under the button (see Figure 3). Note that if the user wants to exclude some files, the option to *Delete* a listed file is available by right-clicking on the undesired entry. Files from different directories can be added by adding one directory at the time with the same procedure.



```
C:\... \DLBC
    ● Mino.vcf.out
    ● MJ.vcf.out
    ● NAMALWA.vcf.out
    ● OCI-LY3.vcf.out
    ● SUP-M2.vcf.out
C:\... \LCLL
    ● KASUMI-2.vcf.out
    ● KOPN-8.vcf.out
    ● KYO-1.vcf.out
    ● MOLT-3.vcf.out
    ● MUTZ-5.vcf.out
```

Figure 3: *Two directories have been added and all the file in each directory are listed in the hierarchy. Details about file format and other requirements are described in the text.*

Finally the user can add a phenotype text file in the D3Oncoprint project, by using the *Browse* button in the lower part of the *Data Import Wizard* tab and select the appropriate file in the pop-up window. This will load the file path in the text field of the GUI. For information about the structure of this file and the accepted formats, please refer to Section 4.3 on page 17.

Once all the fields have been properly populated, the user should click the *Load and validate files* button: D3Oncoprint will scan all the variant files and extract all the available headers. The software will also automatically populate some fields (when possible) and the interface focus will be moved to Tab 2, where the user will select the data columns for the interactive HTML viewer.

Note: if any problem is found while parsing the input data, phenotype file or while writing in the project directory, D3Oncoprint will display an error message and it will try to write more information about the error in a textual log file located in the same output folder as the final HTML. Refer to the message in the log to solve the issue and contact the authors of this document for questions.

3.3.1 Use case example: Colorectal cancer cohort (TCGA-READ)

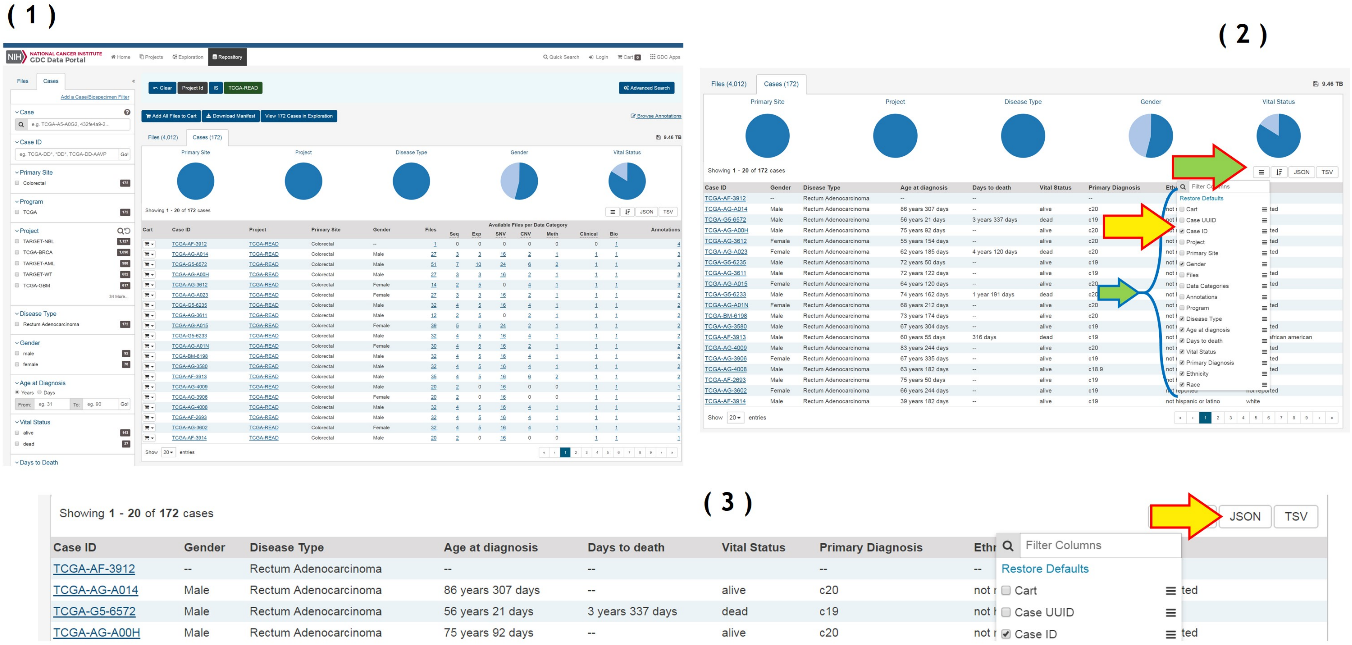
As a user case for this document we are going to use publicly available TCGA variant files of Colorectal cancer (READ). The files are hosted in the GDC data portal (`https://gdc-portal.nci.nih.gov/`) and the specific input file used in our demo is available at the following link (`https://api.gdc.cancer.gov/data/c999f6ca-0b24-4131-bc53-1665948f8e3f`).

A Mutation Annotation Format (MAF) is a tab-delimited file containing somatic mutation annotations. MAF files containing any germline mutation annotations are kept in the controlled access portion of the GDC data portal. In the open access portion of the data portal, MAF files contains only somatic mutations in genes that are not listed in dbSNP. MAF files with different variant call and annotation pipeline are shared in the GDC portal: for the presented demo we use the MAF file from the Mutect pipeline.

A single MAF file from the GDC portal contains the somatic mutations of multiple patients. We need to extract the MAF file from the tar.gz archive and save it in a local directory (called MUTECT in the figures below). In D3Oncoprint MAF files are split into single sample variant file and then processed in the same way as regular tab delimited files. Because of the well-defined and controlled structure of these MAF files, D3Oncoprint is also able to prepopulate some of the mandatory fields presented in Section 3.4 on page 11.

At the time of writing, downloading the phenotype information for the entire dataset of TCGA-READ requires few extra steps. The table is available at the following link `https://portal.gdc.cancer.gov/repository?facetTab=cases&filters=~%28op~%27and~content~%28~%28op~%27in~content~%28field~%27cases.project.project_id~value~%28~%27TCGA-READ%29%29%29%29%29&searchTableTab=cases` (corresponding to the query cases.project.project_id in ["TCGA-READ"]). The displayed table shows only some of the available phenotypes and we want to download all the available data. Figure 4 on the next page guides the user through the steps of downloading the correct phenotype information from the TCGA dynamic table. Using the icon with the three horizontal lines, select all the available columns. Note that it is really important to include the column *Submitter ID* in the table as that column contain the ID that will link the phenotype to the variant files. You can remove columns relative to the number of files or annotations. Clicking the *JSON* button will allow the user to download the entire table in JSON format which is the only format currently supported by the GDC portal for the export of the entire table. When importing this to D3Oncoprint the software automatically recognizes the JSON format and executes a flattening code before generating the interactive HTML page. The generated flat txt file is stored in the same location of the JSON file and we recommend the user to open the generated flat file and manipulate it as appropriate to make headers' labels more human-readable, remove unnecessary columns, modify format of values (e.g. age is expressed in days in GDC JSON files, while it may be more appropriate to convert it into years). Those manipulations can be easily done in any spreadsheet program (e.g. MS Excel). The modified TXT file, saved again as a TAB separated file, can then be used as phenotype information instead of the initial JSON file. At the time of writing, the field that contains the sample/patient id in the JSON file from the GDC portal is called 'submitter_id': in order to be flatten correctly, any JSON file will need to contain a field with that name and its content should be the name of the variant file that phenotype JSON object refers to. If the 'submitter_id' field is not found, D3Oncoprint won't be able to process correctly the JSON file. The 'submitter_id' JSON field correspond to the Case ID column in the web interface, so make sure to select that column before saving the JSON data.

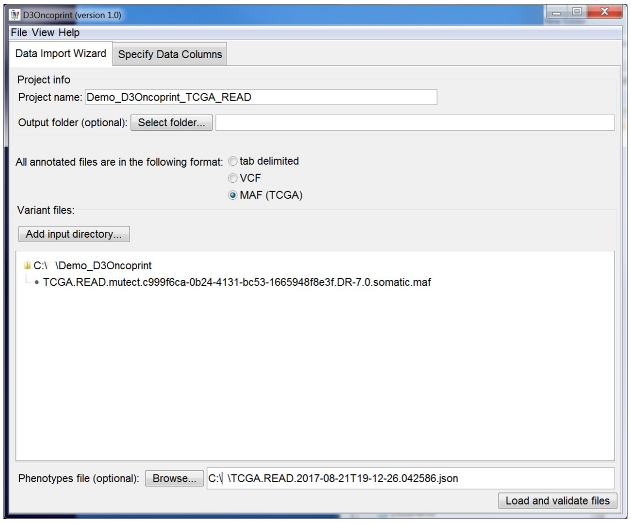
In Figure 5 on the following page we gave a name to the project, selected the directory where we extracted the downloaded MAF file, and selected the downloaded JSON file. At this point we are ready to click the *Load and validate files* button to parse these input files. The following steps will be illustrated in Section 3.4.1 on page 11.



Showing **1** - **20** of **172** cases

Case ID	Gender	Disease Type	Age at diagnosis	Days to death	Vital Status	Primary Diagnosis	Ethi
TCGA-AF-3912	--	Rectum Adenocarcinoma	--	--		--	--
TCGA-AG-A014	Male	Rectum Adenocarcinoma	86 years 307 days	--	alive	c20	not r
TCGA-G5-6572	Male	Rectum Adenocarcinoma	56 years 21 days	3 years 337 days	dead	c19	not l
TCGA-AG-A00H	Male	Rectum Adenocarcinoma	75 years 92 days	--	alive	c20	not r

Figure 4: *Steps to take on the GDC data portal to download the phenotype table for the TCGA-READ cohort. (1) The default content of the table, as displayed when clicking the link provided in the text above. (2) Clicking the icon with the three horizontal bars (yellow arrow at the top) opens a list of all the available headers. Make sure to activate the checkbox of the column Case ID (yellow arrow in the list). Any other additional column of interest can be added by activating the correspondent checkbox. Columns of field that are not relevant can be excluded by emptying the selection of the checkbox. (3) Once the user has completed the selection of the header, click the JSON button and download the table in JSON format.*



Project info

Project name: Demo_D3Oncoprint_TCGA_READ

Output folder (optional):

All annotated files are in the following format: ○ tab delimited ○ VCF ● MAF (TCGA)

Variant files:

C:\ \Demo_D3Oncoprint
- TCGA.READ.mutect.c999f6ca-0b24-4131-bc53-1665948f8e3f.DR-7.0.somatic.maf

Phenotypes file (optional): C:\ \TCGA.READ.2017-08-21T19-12-26.042586.json

Figure 5: *Tab 1, with all the required information for our demo example. A project name has been assigned, the output directory will be the local directory of the jar (so no path is required in the output file textfield), the variant type has been selected as MAF (TCGA), the single MAF file for the TCGA-READ cohort has been loaded in the file tree and the json file containing the phenotype information has been loaded in the phenotypes textfield. Clicking the 'Load and validate files' will check the input file and populate the fields in Tab 2 (see Figure 6 on the next page).*

3.4 Specify Data Columns (Tab 2)

In order to properly parse the information in the original files, D3Oncoprint needs the user to identify which of the available headers contains the gene symbol, which header contains the variant type information and which header contains the aminoacid change information. A dropdown with all the available headers is presented to the user and a selection must be made in order to create a correct interactive viewer.

VCF files have information stored in complex fields (e.g. INFO): D3Oncoprint will separate those fields and present them to the user as single 'headers'. In these situations the variant type column and aminoacid change column could point to the same 'header': D3Oncoprint will try to parse those complex field and extract the required information. For information about how vcf are processed, please refer to Section 4.2 on page 17.

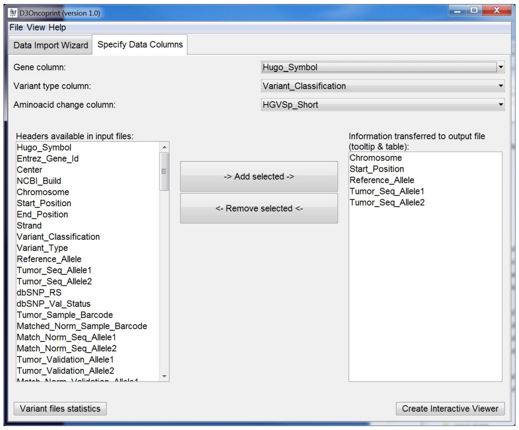
If the information from the three mandatory fields cannot be parsed appropriately for all the files, entries in the variant table may be empty and the d3 map may not be colored correctly. Below the three mandatory field, the user finds a complete list of all the available headers (on the left) and headers of interest can be moved to the right, as the list of the headers that will be used for tooltip and header. Note that D3Oncoprint prepopulate the list with columns that are usually of interest (e.g., chromosome number, start position, alternate base(s)) but the user is free to add/remove any of the available headers.

Including a large number of columns for the tooltip and table may lead to a longer wait to open the interactive HTML page in the web browser and, in some cases, that may result in the page hanging and the third-party libraries to become completely unresponsive. In those situations the best course of action is to close the browser, remove some of the selected columns and re-create the interactive viewer. D3Oncoprint is designed to support the iterative nature of exploring user data, so going back and forth to test different visualizations, included information and grouping is part of how we envisioned our software to be used.

Once the selection is complete, the user can click the *Create Interactive Viewer* button: D3Oncoprint will process each variant file, extract the relevant information and populate the javascript objects in the output HTML page. Upon successful creation of the page, D3Oncoprint will automatically open the page with the system default web browser (we recommend the use of Google Chrome).

3.4.1 Use case example: Colorectal cancer cohort (TCGA-READ)

As mentioned in Section 3.3.1 on page 9, TCGA MAF files are split into single patient variant file and then processed in the same way as regular tab delimited files. Because of the well-defined structure of these MAF files, D3Oncoprint is able to prepopulate some of the mandatory fields in the *Specify Data Columns* tab.



Gene column: Hugo_Symbol

Variant type column: Variant_Classification

Aminoacid change column: HGVSp_Short

Headers available in input files:
Hugo_Symbol
Entrez_Gene_Id
Center
NCBI_Build
Chromosome
Start_Position
End_Position
Strand
Variant_Classification
Variant_Type
Reference_Allele
Tumor_Seq_Allele1
Tumor_Seq_Allele2
dbSNP_RS
dbSNP_Val_Status
Tumor_Sample_Barcode
Matched_Norm_Sample_Barcode
Match_Norm_Seq_Allele1
Match_Norm_Seq_Allele2
Tumor_Validation_Allele1
Tumor_Validation_Allele2
Match_Norm_Validation_Allele1

Information transferred to output file (tooltip & table):
Chromosome
Start_Position
Reference_Allele
Tumor_Seq_Allele1
Tumor_Seq_Allele2

Figure 6: *Tab 2, with all the required information for our demo example. The gene symbol is stored in the 'Hugo_Symbol' column, the variant type is stored in the 'Variant_Classification' and the aminoacid change column is 'HGVSp_Short' (as the short one-letter format of the string is preferred). Default columns added for the datatable presentation and tooltip are 'Chromosome' number, 'Start_Position' and 'Reference_Allele'. Clicking the 'Create Interactive Viewers' will loop through all the files and extract the information to create the HTML page with the final viewer (see Figure 9 on page 16).*

Since the phenotype file was a JSON file, D3Oncoprint created a new phenotype txt file (with the same name of the original file plus the '_generated_D3Oncoprint' postfix). This flatten file allows the user to manipulate the phenotype file to make headers' labels more human-readable, remove unnecessary columns, modify format of values (e.g. age is expressed in days in TCGA JSON files, while it may be more appropriate to convert it into years), etc. Those manipulations can be easily done in any spreadsheet program (e.g. MS Excel). The modified TXT file, saved again as a TAB separated file, can then be used as phenotype information instead of the initial JSON file. The result of using a cleaned phenotype file can be seen in Section 3.5.4 on page 15.

At this point we are ready to click the *Create Interactive Viewer* button to parse all the input files and extract the selected information. The following steps will be illustrated in Section 3.5.4 on page 15.

3.5 Interactive viewer (HTML page)

The interactive viewer created by D3Oncoprint consists of an HTML page with a dynamic D3 oncoprint map and a correspondent datatable containing all non-synonymous variants found in the annotated variant files.

The HTML page is stored in the user selected output directory, together with all the necessary support files (icons, javascript libraries, intermediate files, etc.). D3Oncoprint Java GUI automatically opens the generated HTML page with the system default browser, but – due to the inappropriate support of certain browsers for our dependency javascript libraries – we recommend the use of Google Chrome. If the user doesn't wish to change the operating system default browser, the HTML file can be manually opened with Google Chrome by selecting the proper 'Open with...' option available in any operating system.

All the javascript libraries required by the viewer come bundled in the D3Oncoprint executable jar and the user does not need to download any additional dependency.

For information purposes, the included libraries are DataTables 1.10.12 (`https://datatables.net/download`) with several sub-libraries (JSZip 2.5.0, pdfmake 0.1.18, Buttons 1.2.2, Column visibility 1.2.2, HTML5 export 1.2.2, Print view 1.2.2, ColReorder 1.3.2, FixedColumns 3.2.2, FixedHeader 3.1.2, Responsive 2.1.0, Scroller 1.4.2, Select 1.2.0); jQuery 3.1.0 (`http://jquery.org`); jscolor 2.0.4 (`http://jscolor.com`); jspdf (`https://github.com/MrRio/jsPDF`); spin 2.3.2 (`http://spin.js.org/`); d3 3.5.16 (`https://d3js.org/`).

We worked to optimize the performances of the interactive HTML viewer, however some restrictions on the quality of the generated PDF, on the exported images, or on the speed of the datatable sorting/filtering are outside our control as they depend on the implementation of the included libraries.

Typical plots/tables including one hundred visible genes (d3 rows), one hundred samples/patients (d3 columns), ten phenotype columns, ten columns extracted from the variant files result in just few seconds delay in refreshing/sorting/exporting and high quality PDF or images. Increasing any of those variables may lead to a longer wait and, in some cases, that may result in the page hanging and the third-party library to become unresponsive. The timeout for unresponsive scripts is browser dependent (the default choice is usually around 30 seconds): processing large datasets may take longer than 30 seconds, so before closing the browser or killing the page, consider waiting for it to become responsive again. Waiting time of few minutes are normal for datasets including hundreds of samples, hundreds of variants, several phenotypes and several tooltip columns. In cases where the waiting time exceeds few minutes (or when files are small) the best course of action is to close the browser and re-open again the HTML page: this will cause the page to be reloaded with our default settings and the user can then try to reduce the number of visible genes or select a different options. In order to reduce the number of visible phenotypes or tooltip columns extracted from the variant files, the user will need to reprocess the files from the Java Graphical Interface (as shown in Section 4 on page 17 and Section 3.4 on the previous page): since users can save previously created D3Oncoprint projects, this step should not be too time consuming.

Note: D3Oncoprint is designed to support the iterative nature of exploring user data, so going back and forth to test different visualizations, included information and grouping is part of how we envisioned our software to be used.

Each interactive HTML page is divided into different sections, explained in each separate section below.

3.5.1 *General information and action buttons*

On the left hand-side of the page, the user can find general information about the project like name and information about the number of overall mutations across the samples (see Figure 7 on the following page (1)).

Following the general information, we have a list of expandable buttons that are used to customize the look and content of the D3 heatmap (and connected datatable below). Clicking on some buttons will expand the space right below them and expose nested buttons. An example of the available buttons is shown in Figure 7 on the next page

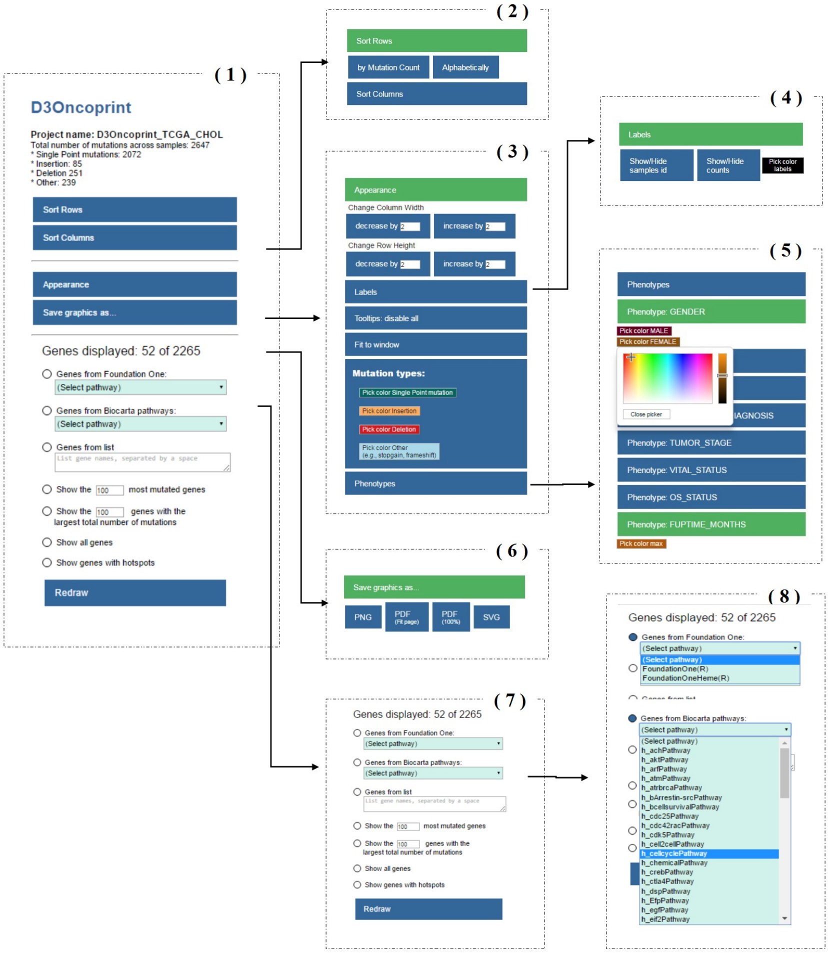


Figure 7: Hierarchy of expandable buttons. (1) General information about the project, name and total number of insertion, deletion and other non-synonimous mutations across all samples. The first level of buttons is also displayed. (2) These buttons control the sorting of rows and columns of the d3 map. (3) Several appearance features of the d3 map can be changed using buttons collapsed in the 'Appearance' button: rows and columns dimension can be modified, tooltips can be enabled/disabled, the entire map can be resized to fit the window and mutation type colors can be modified using an embedded color picker. (4) Labels can also be hidden/displayed and the text color can be selected using the embedded color picker. (5) each phenotype value found in the input file is assigned a different color (randomly) and users can modify each of them clicking on the desired button and picking the new color from the embedded color picker. (6) once the user is happy with the appearance of the d3 map, the graphics can be exported to different image file formats. (7) The content of the d3 map can be controlled by the radio button showed in this part of the window. By default, the 50 most mutated genes across all samples are displayed. However user can query other group of genes, from custom lists (space separated), or from some prepopulated gene lists (e.g., Foundation One, or Biocarta pathways as shown in (8). The user can also ask the viewer to display a certain number of genes, sorted according mutation counts. Genes with curated variants can also be easily accessed using the appropriate option. After the desired option is picked, the user needs to click 'Redraw' to generate the new d3 map)

3.5.2 D3Oncoprint heatmap

Historically, heatmaps are a broadly used tool to explore genomic data because they provide a concise graphical representation of the mutation profile in group of samples. Recently authors of the c-bioportal project (Cerami et al. (2012)) used heatmaps to visually summarize genomic alterations. They developed OncoPrints where individual genes are represented as rows and individual samples are represented as columns. Colored glyphs are then used to visually encode genomic alterations, like somatic mutations. This format can be extremely useful for identifying trends such as mutual exclusivity or co-occurrence between gene pairs. In addition to the web interface, they also provide an R package, ComplexHeatmap (Gu et al. (2016)), which allows researchers to create their own OncoPrints with their own data. Using this package requires advanced programming skills and a complex preprocessing of the variant files as the analysis cannot directly start from commonly used variant format files. Additionally, the generated heatmap is a static image that is difficult to use during the dynamic exploration of information hidden in the genetic mutation profiles.

D3 (Data-Driven Documents) is a JavaScript library for visualizing data using web standards. D3 helps bring data to life using SVG, Canvas and HTML. D3 combines powerful visualization with dynamic interaction capability, making the exploration of usually static heatmaps much more effective and easy, as no programming skills are required from the user, who interacts with the plot just by clicking buttons, dragging labels, moving the mouse cursor to trigger popups, etc.

In D3Oncoprint we wanted to combine the concise representation of OncoPrints, with the dynamic exploration features of D3 objects, and we allow this process to start directly from commonly used variant format files.

Figure 8 shows a typical D3Oncoprint. The caption of the figure contains details on the interactive features of the map. Refer to Section 3.5.4 on the next page for an example of how these features enhance the discoveries of trends in variant data.

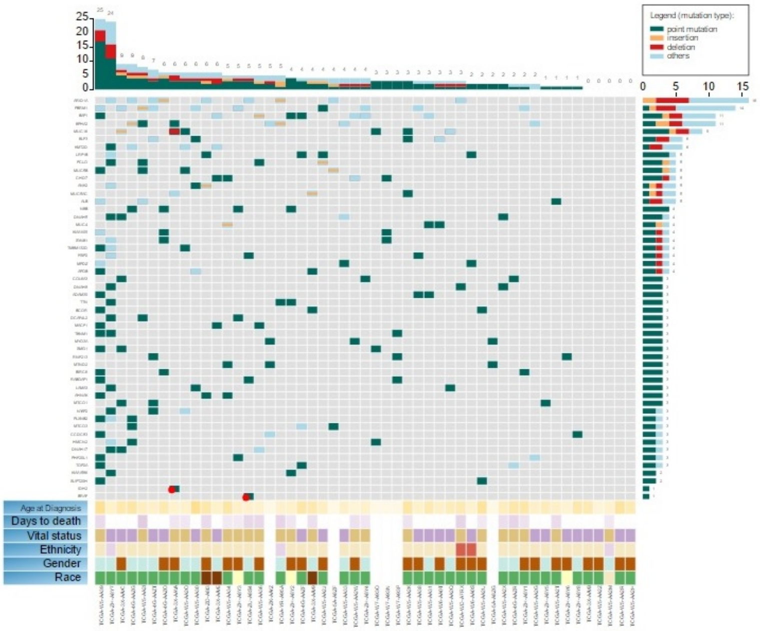


Figure 8: *D3Oncoprint heatmap. Rows represent genes, and columns represent samples. The default view sorts the genes by total mutation count (across all the samples) and the samples by total mutation count (across all the genes). Different types of mutations are identified by different colors (point mutations, insertion, deletions, etc). Moving the cursor on a cell with mutation will display a popup with some information about that mutation. Rows can be dragged up/down by clicking and moving the gene name label. Columns can be dragged left/right by clicking and moving the sample name label. Phenotypes are listed in the lower part of the map. Phenotypes labels are sorting buttons: clicking any of these buttons will sort the columns/samples according to the clicked phenotype value (see Figure 9 on page 16).*

3.5.3 Variant table

In the lower portion of the page, below the action buttons and d3 heatmap, a dynamic table contains all the variants found in all the input files. The table is displayed using the DataTables plug-in, which is a flexible tool that advances interaction controls to any HTML table. Columns can be hidden or showed by accessing the *Column Visibility* button on the top of the table, the table can be exported to commonly used table formats (CSV and Excel). The exported content can be the full table or only the rows compliant with the selected filters. Filter options (textual or numerical) are displayed on top of each column and any content entered in a filter will remove from the view table rows that do not contain the text or fall ouside the numerical range indicated by the filter. The table overall table is designed to quickly load large datasets without too much delay as only few rows are display at any time, while the entire content is stored in the HTML page.

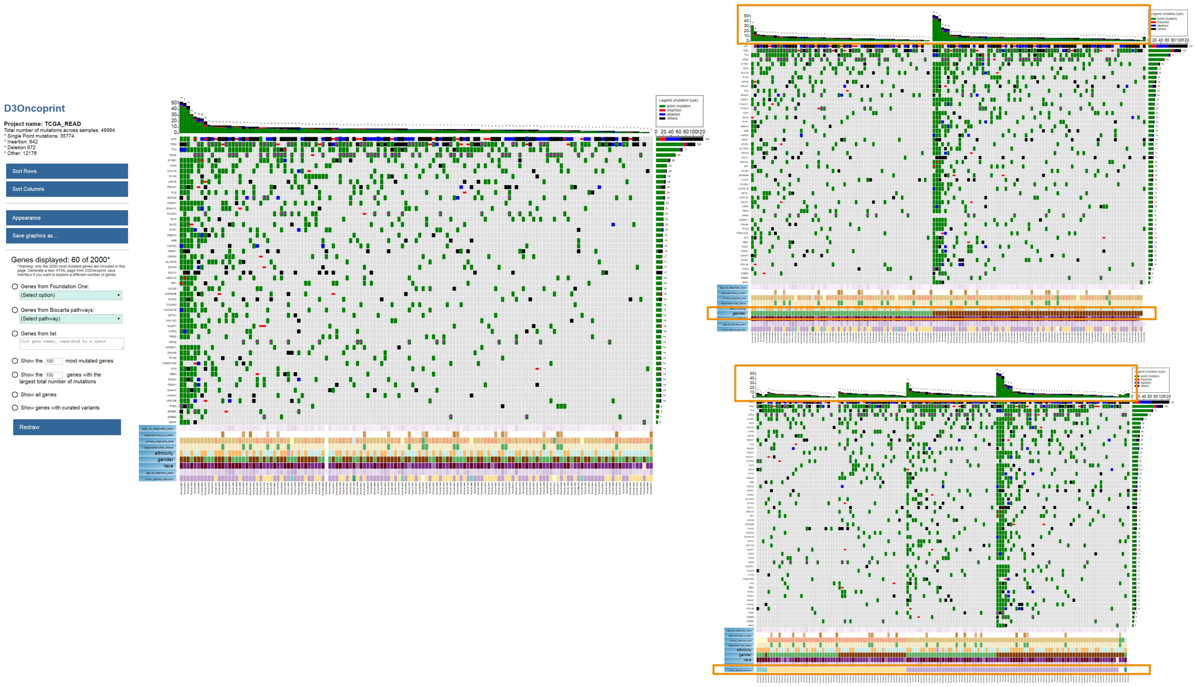
Each row in the table represents a separate variant mutation found in one of the input files. Each row contains the following basic information:

- row number: a numerical id used only for reference purposes. It is a sequential number for the currently displayed content;
- curated variants: icons (and weblinks) marking specific variants as curated variants. For specifics about the hotspot characterization, please refer to Section 4.4;
- sample id: the identifier linking the variant to the phenotypes. The ID comes from the name of the variant file (without extension)
- gene name: gene id for the variant represented in the row. The gene name is a web-link to the GeneCard (`http://www.genecards.org/`) entry for that particular gene;
- variant type: the type of the variant as indicated in the original annotated variant file. Typical values can be ’missense mutation’, ’frameshift’, ’nonsense mutation’, etc
- variant: the specific aminoacid change that define the variant. The variant uses the single letter aminoacid codes;
- *user selected columns*: any column selected in the Java GUI as showed in Section 3.4 on page 11. Default selection include chromosome number, start position and reference allele. If a column contains a cosmic id string (i.e., the string COSM followed by the numerical id of the cosmic entry), this information will be displayed in the table as an weblink to the COSMIC website, so that the user can directly access the relevant information. The number of columns selected affects the performances of the datatable, so we recommend the user to only include columns that are relevant for the analysis and explore different columns in separate iterations of the data exploration;
- *phenotypes*: each phenotype value is appended to each variant row. In this way exporting the table to CSV, the user could perform more in-depth analysis outside our interactive viewer

3.5.4 Use case example: Colorectal cancer cohort (TCGA-READ)

Below we can now explore the TCGA-READ dataset in the generated interactive HTML viewer.

Figure 9 and Figure 10 show the dynamic heatmap and the dynamic datatable. Using buttons and options displayed in drop down menu, we can quickly explore different groups of genes. We can easily sort the dataset according to mutation count or to specific phenotypes. We have also easy access to the complete variant information in the datatable below. The datatable also display links to few curated variants found in our samples: for example we have different KRAS mutations that have been well documented, hence several informational resources exist on My Cancer Genome, CIViC, OncoKB and FDA. External links are displayed as icons and additional information may be present if we move our mouse on these icons.



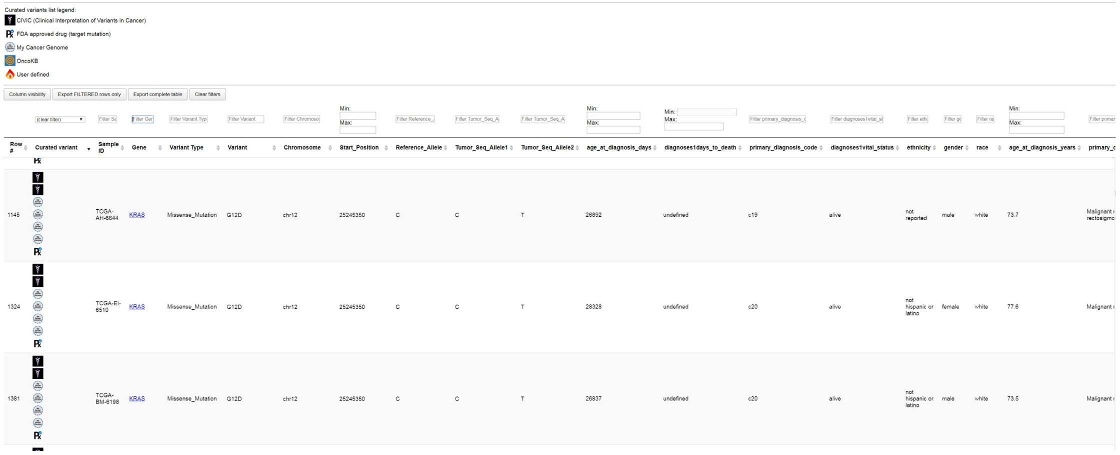
D3Oncoprint

Project name: TCGA_READ
Total number of mutations across samples: 49564
* Single Point mutations: 35774
* Insertion: 642
* Deletion 972
* Other: 12176

Genes displayed: 60 of 2000*
*Warning: only the 2000 most mutated genes are included in this page. Generate a new HTML page from D3Oncoprint Java interface if you want to explore a different number of genes.

○ Genes from Foundation One: (Select option)
○ Genes from Biocarta pathways: (Select pathway)
○ Genes from list: List gene names, separated by a space
○ Show the 100 most mutated genes
○ Show the 100 genes with the largest total number of mutations
○ Show all genes
○ Show genes with curated variants

Figure 9: *D3Oncoprint heatmaps. For display purposes, we decided to reduce the number of phenotypes and changed the labels to make them more readable (compared to the original labels in the json file). On the right we can see the default sorting of the samples (according to the total mutation count). Clicking on the 'Gender' phenotype button will cause a resorting of the columns into two different groups (F, M) as displayed in the top right figure. From the distribution displayed above the heatmap, 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 mutation is all composed of men. If now we click the 'Primary diagnosis code' phenotype button, this will cause a resorting of the columns into multiple groups (c18.9, c19, c20, undefined and c80.1), maintaining the previous Female/Male sorting within each group (Figure in the bottom right). From the distribution displayed above the heatmap, we can see that the group with c20 diagnosis (Malignant neoplasm of rectum) has the highest number of mutations for both genders, and all the male patients with the highest number of mutations have c20 as primary diagnosis. If this result is significant or not will require further analysis of the dataset (especially considering the small sample size) but this is the kind of observations (quickly available using D3Oncoprint) that can guide further investigations.*



Curated variants list legend:
CIViC (Clinical Interpretation of Variants in Cancer)
℞ FDA approved drug (target mutation)
My Cancer Genome
OncoKB
User defined

Row #	Curated variant	Sample ID	Gene	Variant Type	Variant	Chromosome	Start_Position	Reference_Allele	Tumor_Seq_Allele1	Tumor_Seq_Allele2	age_at_diagnosis_days	diagnoses1days_to_death	primary_diagnosis_code	diagnoses1vital_status	ethnicity	gender	race	age_at_diagnosis_years	primary_c
1145	℞	TCGA-AH-6644	KRAS	Missense_Mutation	G12D	chr12	25245350	C	C	T	26892	undefined	c19	alive	not reported	male	white	73.7	Malignant r rectosigmo
1324	℞	TCGA-EI-6510	KRAS	Missense_Mutation	G12D	chr12	25245350	C	C	T	28328	undefined	c20	alive	not hispanic or latino	female	white	77.6	Malignant r
1381	℞	TCGA-BM-6198	KRAS	Missense_Mutation	G12D	chr12	25245350	C	C	T	26837	undefined	c20	alive	not hispanic or latino	male	white	73.5	Malignant r

Figure 10: *D3Oncoprint datatable. The datatable displays links to few curated variants found in our samples: KRAS mutation (with several informational resources from My Cancer Genome, CIViC and FDA showed as icons). External links may have additional information as mouse tooltip. The datatable has sorting, filtering and exporting options.*

4 Details on the input format

In this section we describe the requirements for the input files used by D3Oncoprint. Validation and error checks are included in the code, so if mandatory fields are not appropriately formatted messages would be displayed to the user. Some issues that are not critical for the correct functioning of D3Oncoprint are logged in a textual log file for future reference and for debug purposes. Please refer to these log files (stored in the same output directory as the main project selected by the user) before submitting any bug report to the authors.

4.1 Tab delimited annotated variant files

The most typical annotated variant file output format is TAB delimited. Software and annotation pipelines (like ANNOVAR Wang et al. (2010)) use this as their main format. Each column represents a different kind of information (gene symbol, chromosome number, start, ref, alt, etc.) and D3Oncoprintwill parse all the columns present in the file.

As described in Section 3.4, it is the user responsibility to know which column contains the mandatory information (gene symbol, variant type and aminoacid change).

In D3Oncoprint we want to extract from the annotated variant files only DNA mutations that may play an important role in the study of the disease, so we exclude variant rows that contain any of the following types: ’synonymous’, ’unknown’, ’RNA’, ’Silent’, ’Intron’. Any other type of mutation will be included in the datatable and text filters can be used to only display specific types.

4.1.1 Mutation Annotation Format (MAF)

A Mutation Annotation Format (MAF) is a tab-delimited file containing somatic mutation annotations. MAF is the format used by TCGA to share their dataset. MAF files containing any germline mutation annotations are kept in the controlled access portion of the Data Portal, MAF files containing only somatic mutations are kept in the open access portion of the Data Portal. MAF files with different variant call and annotation pipeline are shared in the GDC portal.

In the first version of D3Oncoprint we split each MAF file into the single sample TAB delimited files before further processing. The column used as a name of the file (and sample id) is ’Tumor_Sample_Barcode’ and it should be included in all MAF files. If the sample/patient ID is stored in a different column, the user can specify the name of the alternative column in the appropriate textfield of the ’File/Advanced option’ menu of D3Oncoprint. The single patient variant files created by this process are then processed in the same way as regular tab delimited files. The files will be removed by D3Oncoprint at the end of the session.

4.2 Variant Calling Format (VCF)

Variant Calling format (Danecek et al. (2011)) is another very popular format for annotated variant files.

Software and annotation pipelines (like snpEFF (Cingolani et al. (2012))) use this as their main format. While some of the information is in a tab delimited format (e.g. chromosome number, start, alt, ref), many of the information needed by D3Oncoprint about the gene symbol, variant type and amino acid change are usually stored in the single INFO column. For VCF files, D3Oncoprint will extract each single ID field within the INFO column and present it in the java interface as a different ’column’.

As described in Section 3.4, it is the user’s responsibility to know which column contains the mandatory information (gene symbol, variant type and amino acid change). If multiple information is saved under a same ’column’ (for example both variant type and aminoacid change could be stored under the same ’ANVR_ANNOT’ id), the user can indicate this, and D3Oncoprint will try to parse and extract the correct information. Note that this process may not cover all possible cases of VCF fields, so we suggest the user to consider options in the annotation pipeline which will favor the splitting of the required information over multiple fields.

In D3Oncoprint we want to extract from the annotated variant files only DNA mutations that may play an important role in the study of the disease, so for VCF file we include only variant rows that contain any of the following types: "nonsynonymous", "SNV", "substitution", "missense", "frameshift", "stop", "start", "insertion", "deletion". Any other type of mutation will NOT be included in the datatable or D3 heatmap.

4.3 Phenotypes

Attaching phenotype information to each sample can help identifying trends and explore the mutation profiles of different subgroups of samples/patients. Typical phenotype data include gender, race, age, histology, survival time, treatment response, etc. Phenotypes can be continuous or discrete, numerical or categorical.

In addition to the group of annotated variant files, D3Oncoprint allows the user to upload a single textual file containing the phenotype information of the batch of samples/patients.

The phenotype file should comply to the following rules:

- must be a TAB delimited text file (typically stored as a .txt file);
- each column represents a different phenotype;
- each row represents a single sample id/patient;
- the first row is the header, and it is used to extract the phenotype names;
- the first column must contain the sample id/patient id, and it must correspond to a name of a variant file uploaded in the interface. The id must correspond to the file name without the extension. Rows not having a matching variant file will be discarded and not used in the HTML display;
- single entries in a row can be empty, if a specific phenotype is not known for a specific sample/patient;
- D3Oncoprint will automatically parse the phenotype values and determine if the variable is a number or a string. When a number is found, if more than 5 distinct values are found, that phenotype will be considered continuous; if the value is a string or a number with less than 5 distinct values, the phenotype will be considered discrete.

To add the phenotype file in the D3Oncoprint project, use the *Browse* button in the lower part of the *Data Import Wizard* tab and select the appropriate file in the pop-up window (Figure 5 on page 10). This will load the file path in the text field of the GUI.

When the user clicks the button to generate the interactive HTML viewer (after the selection of the necessary columns as explained in Section 3.4 on page 11), D3Oncoprint will parse the information in all the annotated variant files and, combining it with the information in the phenotype file, the interactive HTML page with the D3 oncoprint map will be generated. For an explanation of all the dynamic exploration features related to the phenotype labels, please refer to Section 3.5.2 on page 13. An example of the powerful data exploration that D3Oncoprint facilitate is shown in Figure 6 on page 11.

D3Oncoprint accepts also JSON format for phenotype files (`http://json.org/`). JSON is an open-standard format that uses human-readable text to transmit data objects consisting of attribute-value pairs. At the time of writing, this is the only format supported for the phenotype table on the GDC portal for TCGA data (see Figure 5 on page 10). Since JSON has an arbitrary complex hierarchy of nested elements D3Oncoprint processes a JSON file and flatten it to a TAB delimited txt compliant with the rules described above. D3Oncoprint automatically recognizes if a JSON file is selected in the phenotype textfield and executes the flattening code before generating the interactive HTML page. The generated flat txt file is stored in the same location of the JSON file and we recommend the user to open the generated flat file and manipulate it as appropriate to make headers' labels more human-readable, remove unnecessary columns, modify format of values (e.g. age is expressed in days in TCGA JSON files, while it may be more appropriate to convert it into years). Those manipulations can be easily done in any spreadsheet program (e.g. MS Excel). The modified TXT file, saved again as a TAB separated file, can then be used as phenotype information instead of the initial JSON file. At the time of writing, the field that contains the sample/patient id in the JSON file from the GDC portal is called 'submitter_id': in order to be flatten correctly, any JSON file will need to contain a field with that name and its content should be the name of the variant file that phenotype JSON object refers to. If the 'submitter_id' field is not found, D3Oncoprint won't be able to process correctly the JSON file.

4.4 Curated variants lists and other support files

In addition to display the information contained in the input variant files, D3Oncoprint includes curated information used to facilitate the discovery of meaningful biological information hidden in the genomic data.

The curated information included in the first version of the software include:

- Gene lists, grouped according to specific functions (e.g. BioCarta pathways (`https://cgap.nci.nih.gov/Pathways/BioCarta_Pathways`), FoundationOne genes (`http://foundationone.com/`)).
  These lists are made available in the HTML page, and they are used to quickly display only the mutated genes that belong to a specific group;
- Hotspot mutations:

  CIViC These are variants of interest in CIViC (an open access, community-driven web resource for Clinical Interpretation of Variants in Cancer, Griffith et al. (2016)) (`https://civic.genome.wustl.edu/#/home`)

If a variant in a sample is matched with our curated list from CIViC, an icon will appear next to the entry and a link to the source information is provided for further investigation;

MCG These are variants of interest from My Cancer Genome (an up-to-date resource on what mutations make cancers grow and related therapeutic implications, including available clinical trials., Swanton (2012)) (https://www.mycancergenome.org/).
If a variant in a sample is matched with our curated list from My Cancer Genome, an icon will appear next to the entry and a link to the source information is provided for further investigation;

FDA These are variants of interest in FDA approved drugs for target mutations http://www.fda.gov/Drugs/InformationOnDrugs/ApprovedDrugs/.
If a variant in a sample is matched with our curated list from FDA approved drugs, an icon will appear next to the entry and a link to the source information is provided for further investigation;

OncoKB These are variants listed as “Actionable variants” in OncoKB, a precision oncology resource that contains information about the effects and treatment implications of specific cancer gene alterations Chakravarty et al. (2017);

User Defined This file is completely customizable by users, so that variants of interest for a specific project can be easily marked and highlighted in the maps and tables. The txt file containing this information is stored in the ’.supportFiles’ directory next to the executable JAR of D3Oncoprint and the name of the curated variant file file is ’hotspots_userdefined.txt’. The content can be changed using any text editor, but neither the name, nor the location nor the format of the file should be changed. The file is TAB delimited and it contains four columns (gene, variant_aachange, url, tooltip_info). Each variant of interest should populate at least the ’gene’ and ’variant_aachange’ column. The other two columns are optional but they provide a way to get a reference link and some extra tooltip information embedded in the HTML viewer.

D3Oncoprint authors plan to maintain and update these curated support files, to give some guidance in the exploration of the information hidden in genomic data. The updated information is distributed seamlessly to any D3Oncoprint user connected to the web as the software accesses BRP server each time it is launched. If new versions are found, the user can get the most updated versions with just one click.

References

Cerami, E., Gao, J., Dogrusoz, U., Gross, B. E., Sumer, S. O., Aksoy, B. A., Jacobsen, A., Byrne, C. J., Heuer, M. L., Larsson, E., Antipin, Y., Reva, B., Goldberg, A. P., Sander, C., and Schultz, N. (2012). The cBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data. *Cancer Discovery*, 2(5):401–404.

Chakravarty, D., Gao, J., Phillips, S., Kundra, R., Zhang, H., Wang, J., Rudolph, J. E., Yaeger, R., Soumerai, T., Nissan, M. H., et al. (2017). Oncokb: a precision oncology knowledge base. *JCO Precision Oncology*, 1:1–16.

Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., Land, S. J., Lu, X., and Ruden, D. M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, snpeff: Snps in the genome of drosophila melanogaster strain w1118; iso-2; iso-3. *Fly*, 6(2):80–92.

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., et al. (2011). The variant call format and vcftools. *Bioinformatics*, 27(15):2156–2158.

Griffith, M., Spies, N. C., Krysiak, K., Coffman, A. C., McMichael, J. F., Ainscough, B. J., Rieke, D. T., Danos, A. M., Kujan, L., Ramirez, C. A., et al. (2016). Civic: A knowledgebase for expert-crowdsourcing the clinical interpretation of variants in cancer. *bioRxiv*, page 072892.

Gu, Z., Eils, R., and Schlesner, M. (2016). Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics*, 32(18):2847–2849.

Swanton, C. (2012). My cancer genome: a unified genomics and clinical trial portal. *The Lancet Oncology*, 13(7):668–669.

Wang, K., Li, M., and Hakonarson, H. (2010). Annovar: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic acids research*, 38(16):e164–e164.