Chromatic: WebAssembly-Based Cancer Genome Viewer

Richard Finney and Daoud Meerzaman

Computational Genomics Research Group, Center for Biomedical Informatics and Information Technology, National Cancer Institute, Bethesda, MD, USA.

Cancer Informatics
Volume 17: 1–2

DOI: 10.1177/1176935118771972

ABSTRACT: Chromatic is a novel web-browser tool that enables researchers to visually inspect genomic variations identified through next-generation sequencing of cancer data sets to determine whether such calls are, in fact, valid. It is the first cancer bioinformatics tool developed using WebAssembly technology, which comprises a portable, low-level byte code format that provides for the rapid execution of programs within supported web browsers. It has been designed expressly for ease of use by scientists without extensive expertise in bioinformatics.

KEYWORDS: WebAssembly, cancer, mutation, viewer

RECEIVED: December 26, 2017. **ACCEPTED:** March 21, 2018.

TYPE: Software or Database Review

FUNDING: The author(s) received no financial support for the research, authorship, and/or publication of this article.

DECLARATION OF CONFLICTING INTERESTS: The author(s) declared the following potential conflicts of interest with respect to the research, authorship, and/or publication of this article: This work was performed by employees of the National Institutes of Health.

CORRESPONDING AUTHOR: Richard Finney, Computational Genomics Research Group, Center for Biomedical Informatics and Information Technology, National Cancer Institute, 9609 Medical Center Drive, Bethesda, MD 20892, USA. Email: finneyr@nih.gov

Introduction

High throughput next-generation sequencing (NGS) technologies have provided researchers with the ability to identify millions of genomic variations, but they are not without important weaknesses that can undermine the validity of the results obtained from any study. Despite the use of automated filtering approaches to identify erroneous variants, many remain problematic and require visual inspection, which remains the best approach to differentiate between real variants and artifacts that may include false positives and false negatives. Several visualization tools are available. However, these generally are time-consuming and difficult to use, especially for scientists lacking expertise in using bioinformatics tools.

Here, we report the development and release of an innovative tool, Chromatic, that provides users with a robust capability to observe variants simply and quickly using a web browser. Chromatic is written in portable C which is compiled into JavaScript using Emscripten. Emscripten uses a C compiler which creates LLVM[1] (low-level virtual machine) byte code which is transpiled ultimately into JavaScript. No executable "EXE" file, or native binary, is created.

WebAssembly,[2] a byte code that executes in supported browsers independent of browser add-ons, is a new technology that is much faster than JavaScript, although not as fast as native binary executables. Current implementations of WebAssembly cannot manipulate the browser's Document Object Model (DOM) and rely on interfacing with JavaScript to update the screen. Code originally written for native targets can be ported to the browser without the added difficulties of adapting to different operating systems. Installation is as simple as loading a webpage. Security is the same as any other webpage within the sandboxed environment of the browser.

Features and Methods

Chromatic is designed to allow users to review variant calls including mutations and small insertions or deletions (indels). It provides access to public cancer data sets via a webserver at the National Cancer Institute (NCI). The initial release of Chromatic provides access to 4 publicly available cancer projects: various samples from the Texas Cancer Research Biobank (TCRB),[3] whole genome data for liver samples from the Korean Genome Research Foundation,[4] whole genome liver samples from Beijing Genomics Institute (BGI-Shenzhen),[5,6] and esophageal squamous cell carcinoma exome data from Fudan University.[7] The TCRB data were downloaded from their website. The other 3 data sets were downloaded from the public section of the National Center for Biotechnology's Short Read Archive (SRA). All data were processed using novoalign (http://novocraft.com) on the computational resources of the high performance computing center and the national institutes of health (NIH HPC) Biowulf cluster (http://hpc.nih.gov). Chromatic also provide access to protected data from The Cancer Genome Atlas (TCGA) after users acquire a secure token from the NCI's Genomics Data Commons (GDC).[7]

Users can operate the program using simple menus and dashboard interfaces to view images of genomic regions and navigate through tens of thousands of cancer samples. A batch-mode feature provides for scripting for long-running operations, which creates a slideshow that can be saved to the user's local storage. Batch-mode jobs are stored in tape archive (TAR) files and can be extracted for viewing using a web browser. The resulting slideshow HTML file and PNG image files can help in reviewing a large number of genomic locations, substantially reducing the workload involved in examining many variant calls.

Chromatic source code is available on the website, https://chromatic.nci.nih.gov/chromatic.html. Chromatic uses some of the code for image processing and binary sequence alignment map (BAM) processing from the Alview BAM file viewing project.[8] The functionality for reading BAM files is a custom implementation using the BAM file specification,[9] which has been refined to reproduce the quirks of the samtools

library. Open and lightweight PNG image processing, ZLIB[10] compression code, and TAR code were used from various sources and were compiled and linked as regular C files instead of linked libraries. The purposes of integrating these third-party files in line are to simplify compilation and facilitate porting the code to other systems.

Most of the code created for Chromatic, including the core functionality, is available in the public domain. Chromatic uses C source code files to implement TAR archives, gzip decompression, and PNG image manipulation, all of which have permissive Massachusetts Institute of Technology (MIT)-style licenses (ie, no restrictions other than maintaining attributions exist). Developers may modify the code and use it as they wish.

Chromatic communicates with 2 support programs on the host server from which it is downloaded. One program, called "slicer," is a simple proxy that obtains a subset of the BAM file from the GDC website (https://gdc.cancer.gov) or server local storage. The second support program, called "srvdna," is a common gateway interface (CGI) program[11] that provides selected portions of the human genome reference sequence, thereby allowing researchers to avoid spending the large amount of time it would take to download the entire sequence. The source codes of "slicer" and "srvdna" are provided as part of the base source release of Chromatic for those who want to customize and host their own instance.

Limiting server operations to simple proxy duty and serving up sequence data simplify the role of the server. As the BAM file and sequence are processed within the user's browser, the burdens of complex session management and data processing on the server are eliminated. Moving the work to the browser also significantly simplifies security on the server as it only needs to serve data and proxy requests for short read data. A disadvantage of Chromatic is that there is more network traffic generated than would be in the case of a server program generating the images and serving them up. Also, a true native binary version would run faster than a WebAssembly version.

Summary

Chromatic is an easy to use web-browser tool that provides fast, near-native central processing unit (CPU) performance. WebAssembly bypasses the often onerous installation difficulties associated with obtaining and setting up a native program binary executable. Program invocation is reduced to opening a webpage. The program provides easy access to short read data regardless of which operating system a given user runs.

Developers may reengineer and distribute Chromatic as long as they maintain the attribution notices present in portions of the source code.

Acknowledgements

We thank Dr Laura K. Fleming, our scientific writer and editor in the National Cancer Institute (NCI) Center for Biomedical Informatics and Information Technology, for her critical reading of the manuscript. Availability—Web: https:///chromatic.nci.nih.gov; Source: https://chromatic.nci.nih.gov/chromatic-source.html.

Author Contributions

RF programmed Chromatic and helped write the manuscript. DM managed project and helped write manuscript.

REFERENCES

1. Lattner C, Adve V. LLVM: a compilation framework for lifelong program analysis and transformation. In: Proceedings of the International Symposium on Code Generation and Optimization; March, 2004; San Jose, CA.
2. Haas A, Rossberg A, Schuff DL, et al. Bringing the web up to speed with WebAssembly. In: Proceedings of the 38th ACM SIGPLAN Conference on Programming Language Design and Implementation; June 18-23, 2017:185–200; Barcelona. New York, NY: ACM. doi:10.1145/3140587.3062363.
3. Becnel LB, Pereira S, Drummond JA, et al. An open access pilot freely sharing cancer genomic data from participants in Texas. *Sci Data*. 2016;3:160010.
4. Lee YS, Kim BH, Kim BC, et al. SLC15A2 genomic variation is associated with the extraordinary response of sorafenib treatment: whole-genome analysis in patients with hepatocellular carcinoma. *Oncotarget*. 2015;6:16449–16460.
5. Sung W-K, Zheng H, Li S, et al. Genome-wide survey of recurrent HBV integration in hepatocellular carcinoma. *Nat Genet*. 2012;44:765–769. doi:10.1038/ng.22952.
6. Grossman RL, Heath AP, Ferretti V, et al. Toward a shared vision for cancer genomic data. *N Engl J Med*. 2016;375:1109–1112.
7. Deng J, Chen H, Zhou D, et al. Comparative genomic analysis of esophageal squamous cell carcinoma between Asian and Caucasian patient populations. *Nat Commun*. 2017;8:1533.
8. Finney RP, Chen Q-R, Nguyen CV, et al. Alview: portable software for viewing sequence reads in BAM formatted files. *Cancer Inform*. 2015;14:105–107.
9. Li H, Handsaker B, Wysoker A, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. 2009;25:2078–2079.
10. Deutsch P, Gailly JL. RFC1950: ZLIB compressed data format specification version 3.3. RFC Editor, 1996.
11. Robinson D, Coar KAL. The common gateway interface (CGI) version 1.1. *RFC*. 2004;3875:1–36.