GenomeViewer

An Interactive Genomic Somatic Mutation Visualizer.

Beatriz Kanzki, Alain April
École de Technologie Supérieure (ÉTS)
1100, rue Notre-Dame ouest,
Montréal, QC, Canada
+15 148 855 744
beatriz.kanzki.1@ens.etsmtl.ca
alain.april@etsmtl.ca

ABSTRACT

New Generation Sequencing (NGS) technologies offer new insights to researchers in the field of oncogenomics. These technologies provide valuable genetic information by rapidly detecting and identifying expected mutations to improve clinical treatments. To be used effectively, this large amount of data has to be processed, explored and interpreted carefully and quickly. Meanwhile, cancer research continues to publish new theories and findings based on large-scale collaborative projects that provide publicly available genomic and clinical cancer data. However, researchers have a hard time using the data to its full potential although it's readily available. Between the growing output size and complexity of NGS technologies, and the growing number of publicly available heterogeneous databases, processing and exploring this data can become a challenge for the average researcher. This paper presents GenomeViewer's functionalities, which specializes in visualization of somatic mutations in cancer genomics. This easy to use software will enable cancer researchers to seamlessly compare their data against publicly available resources. GenomeViewer uses "Big Data" technologies such as Spark and Parquet, and is based on the UC Berkeley's Analysis Data Model (ADAM) genomic format for cloud scale computing. Our hope is that GenomeViewer will become the preferred tool for viewing somatic mutations for researchers in cancer genomics.

Keywords

Bioinformatics; GOAT; GenomeViewer; Genomic loci; Somatic Mutations; Indels; Substitutions; Visualization tool, Software Engineering, Spark, Big Data; ADAM;

1. INTRODUCTION

The advent of revolutionary technologies such as Next Generation Sequencing (NGS) has provided new ways and scales that yield new questions for advancing knowledge. From electron microscopy, cell culture and PCR (Polymerase Chain Reaction), NGS is changing the way we understand molecular biological processes [1], and can now provide clinicians with insights that will help them individualize patient care.

First of all, NGS technologies are used to sequence DNA in order to identify mutations or alterations that are acquired or inherited. Sanger's sequencing method has been improved and accelerated with NGS technologies. What used to take months, now takes days; which means that the information provided from an individual's DNA can now be used as a discovery and diagnostics tool for clinicians [2]. In the field of cancer genomics, this technology is used to identify changes during malignant progression, evolution of cancers, and detect genetic variations that predispose an individual to the disease. Furthermore, a number of web-based portals have been created to facilitate access to oncogenomic datasets and assist with their interpretation [3].

However, NGS technologies produce massive amounts of data requiring powerful computational infrastructure, high quality bioinformatics software and skilled personnel to operate them [4]. In addition, although web portals such as The Cancer Genome Atlas (TCGA) provide datasets to the scientific community, they are useless for a major part of the research community and their scientific potential cannot be fully explored [5]. Between the growing amount of data yielded by NGS technologies and the publicly available databases, researchers have difficulty exploring and visualizing all this data. Furthermore, few web-based tools provide means to researchers to compare their own data against that which is publicly available.

To address these issues we have created GenomeViewer, which specializes in the visualization of somatic mutations. First we'll present some of its key features, then we'll present its internal infrastructure.

2. GENOMEVIEWER'S KEY FUNCTIONALITIES

A key feature of GenomeViewer is that researcher's will be able to compare their own data against publicly available genetic information online. In this first version, the functionalities detailed here will be provided:

- Users will be offered the ability to search the genome and compare their data with two options.

  File uploading: where column headers will be detected and will prompt users to provide content of columns by selecting variant identifiers, chromosomes, positions, and alleles.

  Direct typing: User provides variant identifiers or gene name.

- All matches found will be displayed with all possible cancer types or tissue where mutation was found for further selection.
- Databases used to provide genome annotation will be 1000Genome with human genome version 19 (HG19) and HG38 preloaded [3].
- Types of mutations to visualize can be selected and include: substitutions, indels.
- All images will be interactive, and responsive to filters and types of cancer selected, by highlighting the researcher's file for matched biomarkers.

DH '17, July 02-05, 2017, London, United Kingdom

ACM ISBN 978-1-4503-5249-9/17/07.
http://dx.doi.org/10.1145/3079452.3079477

More databases, and functionalities will be added to this tool, as it's hoped that it will satisfy researchers in oncogenomics, and possibly clinicians in identifying mutations and possible targets for treatments.

3. INTERNAL STRUCTURE.

To create such software, a certain amount of quality attributes had to be selected at the conception stage of the project. They included: performance, maintenance, usability, robustness, liability, and testability.

To address testability, and maintenance, a layered architecture was proposed and can be found in figure 1 [4]. This architecture will enable modular development by separating each layers responsibility, and can be tested separately.

To address performance, robustness and liability, we used ADAM format [5], which uses Spark as a compute engine and Parquet for data access. While Spark is an in-memory MapReduce framework which minimizes input and output accesses, Parquet is an open-source store designed for distribution across multiple computers with high compression. This kind of technology has been designed to improve performance, extraction and loading of data by 50X speedup on a 100 node-computing cluster [5] [6]. And extraction of biomarkers in 1.5TB of data was performed at a speed of ~3 seconds versus 5 min with MySQL.

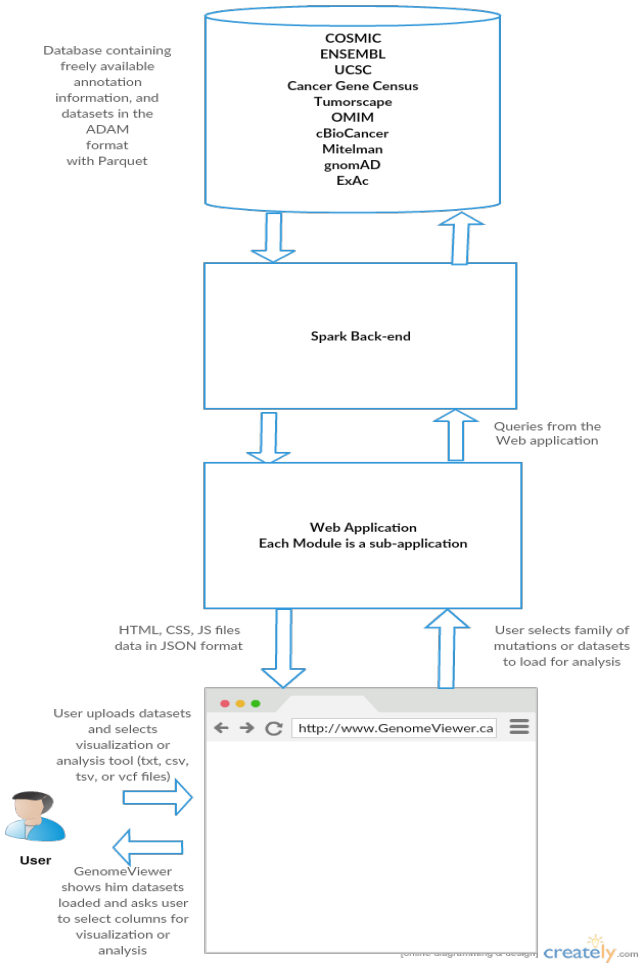

Figure 1. GenomeViewers layered architecture.

To test for usability, tools such as LocusZoom [7], the Integrative Genomics Viewer (IGV) [8], UCSC Genome Browser [9] and the generic genome browser on COSMIC [10], used by researchers in this field have been evaluated. Furthermore, two research groups: the Centre de Recherche du Centre Hospitalier de l'Université de Montréal (CRCHUM), and the Centre de Recherche du Centre Hospitalier Sainte Justine (CRCHUSJ) have voiced that actual tools are time consuming, and do not meet the needs in the field of cancer genomics. This is why, it was decided that GenomeViewer would be specialized in visualization of somatic mutations, and would be built on top of big data to improve performance. Most functionalities useful for this particular field has been kept. Such as visualization of somatic mutations, browsing through the genome and detection of substitutions and indels. What really makes GenomeViewer stand out from other tools is the fact that not only it will be possible for researchers to visualize functional annotations, but it will enable them to compare their data against publicly available databases, interactively as stated in section 2, and without current difficulties linked to performance or amounts of data.

4. CONCLUSION

For this first version, we plan to deploy GenomeViewer on Amazon for distributed computing and to accelerate queries. We hope that GenomeViewer will meet researchers' needs in oncogenomics by offering users a whole new visualization experience. And we hope that it will become a popular tool for visualization of somatic mutations, and that clinicians will find in GenomeViewer an unavoidable companion in specialized medicine, and individualized patient care.

7. REFERENCES

[1] Radford AD, Chapman D, Dixon L, Chantrey J, Darby AC, Hall N. Application of next generation sequencing technologies in virology. Journal of General Virology. 2012 Sep 1; 93 (9):1853-68.

[2] Morozova O, Marra MA. Applications of next generation sequencing technologies in functional genomics. Genomics. 2008 Nov 30; 92 (5):255-64.

[3] Hong MK, Macintyre G, Wedge DC, Van Loo P, Patel K, Lunke S, Alexandrov LB, Sloggett C, Cmero M, Marass F, Tsui D. Tracking the origins and drivers of subclonal metastatic expansion in prostate cancer. Nature communications. 2015 Apr 1; 6.

[4] Clements PC. *Software architecture in practice* (Doctoral dissertation, Software Engineering Institute).

[5] Massie M, Nothaft F, Hartl C, Kozanitis C, Schumacher A, Joseph AD, Patterson DA. Adam: Genomics formats and processing patterns for cloud scale computing. University of California, Berkeley Technical Report, No. UCB/EECS-2013. 2013 Dec 15; 207.

[6] Nothaft FA, Massie M, Danford T, Zhang Z, Laserson U, Yeksigian C, Kottalam J, Ahuja A, Hammerbacher J, Linderman M, Franklin MJ. Rethinking data-intensive science using scalable analytics systems. InProceedings of the 2015 ACM SIGMOD International Conference on Management of Data 2015 May 27 (pp. 631–646). ACM.

[7] Pruim RJ, Welch RP, Sanna S, Teslovich TM, Chines PS, Gliedt TP, Boehnke M, Abecasis GR, Willer CJ. LocusZoom: regional visualization of genome-wide association scan results. Bioinformatics. 2010 Sep 15; 26 (18):2336-7.

[8] Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. Integrative genomics viewer. Nature biotechnology. 2011 Jan 1; 29 (1):24-6.

[9] Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. The human genome browser at UCSC. Genome research. 2002 Jun 1; 12 (6):996-1006.

[10] Forbes SA, Bindal N, Bamford S, Cole C, Kok CY, Beare D, Jia M, Shepherd R, Leung K, Menzies A, Teague JW. COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. Nucleic acids research. 2010 Oct 15:gkq929.