



myGenomeBrowser: building and sharing your own genome browser

Sébastien Carrere, Jerome Gouzy

► To cite this version:

Sébastien Carrere, Jerome Gouzy. myGenomeBrowser: building and sharing your own genome browser. *Bioinformatics*, 2017, 33 (8), pp.1255-1257. 10.1093/bioinformatics/btw800 . hal-01608904

HAL Id: hal-01608904

<https://hal.science/hal-01608904>

Submitted on 25 May 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution - NonCommercial 4.0 International License

Data and text mining

myGenomeBrowser: building and sharing your own genome browser

Sébastien Carrere* and Jérôme Gouzy

LIPM, Université de Toulouse, INRA, CNRS, Castanet-Tolosan, France

*To whom correspondence should be addressed.

Associate Editor: Inanc Birol

Received on September 5, 2016; revised on November 23, 2016; editorial decision on December 9, 2016; accepted on December 13, 2016

Abstract

myGenomeBrowser is a web-based environment that provides biologists with a way to build, query and share their genome browsers. This tool, that builds on JBrowse, is designed to give users more autonomy while simplifying and minimizing intervention from system administrators. We have extended genome browser basic features to allow users to query, analyze and share their data.

Availability and implementation: myGenomeBrowser is freely available at <https://bbirc-pipelines.toulouse.inra.fr/myGenomeBrowser> and includes tutorial screencasts. Source code and installation instructions can be found at <https://framagit.org/BBIRC/myGenomeBrowser>. myGenomeBrowser is open-source and mainly implemented in Perl, JavaScript, Apache and Docker.

Contact: sebastien.carrere@inra.fr

1 Introduction

Today, biologists are able to analyse large volumes of sequencing data thanks to user-friendly academic [Galaxy (Goecks *et al.*, 2010), SMRTPortal] or commercial (CLCbio®, Genious®) solutions. Most projects generate dozens of results files like genome assemblies, structural and functional annotations, read mapping files, variant calling and expression measures.

However, there is still a lack of solutions to visualize these results, especially for integrating and sharing multiple sources of data produced throughout the whole project lifecycle.

In order to handle this task, biologists have two alternatives. The first is to install standalone tools [IGV (Robinson *et al.*, 2011), IGB (Freese *et al.*, 2016), Geneious®, JBrowse-Desktop (Buels *et al.*, 2016)]. Most of the time these solutions fully meet biologists' primary needs. Such tools offer a rich graphical environment to represent results in standard formats. Unfortunately, these tools are limited in terms of data sharing and collaborative operations. The second solution is to ask a bioinformatics platform to make data available via tools implemented on a server and shared via the web [UCSC (Karolchik *et al.*, 2011), GBrowse (Donlin, 2009), JBrowse (Buels *et al.*, 2016), Tripal (Ficklin *et al.*, 2011)]. The disadvantage of this option is that it requires the intervention of an administrator to upload data and implement

authentication solutions to adapt access to different data sources according to user needs.

However, two solutions, combining the advantages of both these alternatives, have been developed in recent years. Trackster (Goecks *et al.*, 2012) provides a way to visualize omics analysis results produced via Galaxy. It therefore benefits from all groupware functionality offered by this environment, such as the system to share and publish a workspace. It also provides the possibility of running analytic software installed in the Galaxy instance on selected genomic regions. But this strong link with Galaxy also has the disadvantage of forcing system administrators to install and maintain the full package even if their users use other software to analyze their data. The second solution proposed by WebGBrowse (Podicheti *et al.*, 2009) is to automatically deploy genome browser from annotation files. However, this solution only allows users to view a single file type (GFF3), which is no longer sufficient as BAM, VCF, BED and bigWig formats are now also widely used as standards. In addition, the tool does not offer built-in data privacy, therefore requiring the system to be reinforced in order to add a layer of authentication.

To cope with biologists' growing need for autonomy as they process their data, we have developed myGenomeBrowser. The

Table 1. Comparison of functional features

	WebGBrowse	Trackster	myGenomeBrowser
Keyword search	—	—	++ ^a
Sequence extraction	+	— ^b	++
Sequence analysis	—	+++ ^c	+ ^d
Share	+ ^e	+++ ^f	++ ^g
Revoke	—	— ^h	++
Input file formats	— ⁱ	+++ ^j	++ ^k
Authentication	— ^l	+ ^m	++ ⁿ

^aAnnotation, alignment and InterPro tracks are searchable,

^bNeeds a dedicated tool to be available in Galaxy,

^cInject data into Galaxy tools and workflows,

^dBlast service,

^eKind of permalink,

^fNominative full access by copy, data can be edited and shared back again,

^gNominative read-only grant access,

^hOnce copied in the shared history, you cannot remotely delete this data,

ⁱOnly gff3,

^jgff3, vcf, bam, bed, bigwig, gtf, wig, bigBed,

^kgff3, vcf, bam, bed, bigwig,

^lwebserver level,

^mGalaxy solution,

ⁿhtpasswd, LDAP, Shibboleth.

software is a web-based environment based on JBrowse and supplemented with various features that are missing from current solutions, such as management of various data sources, mining tools and sharing systems.

2 Results

2.1 Simplifying system administration

myGenomeBrowser is provided with an installation script based on the Docker container management system (Merkel, 2014). Configuration is limited to setting environment variables to define the directory that contains user data. This automatic, simplified configuration allows a bioinformatics platform to quickly deploy an instance of myGenomeBrowser. The default authentication system is based on an 'htpasswd' Apache file. User accounts can be created at shell level by the administrator or online by following a hyperlink received by email at the user's request. For platforms providing a LDAP or Shibboleth-based authentication, a so-called 'expert' configuration protocol is provided. To monitor the use of myGenomeBrowser, the administrator interface provides access to a statistics page that lists all users and the number and disk space occupied by their browsers. The administrator's role is then limited to managing accounts, data management being the responsibility of biologists.

2.2 Visualization

Once the reference sequences have been loaded into the system by the user, that person can add perennial tracks, stored on the server side. myGenomeBrowser can display the various results (annotations, alignments, quantitative data, polymorphisms) produced in the standard GFF3, BAM, BED, bigWig and VCF formats and BLAST tabulated output. The software checks data integrity and consistency before adding these new tracks to the genome browser instance without any additional user configuration. Track rendering is ensured via custom JBrowse configuration templates in order to modify the color of the different biological objects or to identify the impact of polymorphic sites on the reference sequence. The default

configuration offers context menus to extract sequences of biological objects and visualize the InterPro protein domain content (Jones *et al.*, 2014) when available using BioJS (Gomez *et al.*, 2013) components.

2.3 Result mining

The upload of reference sequences and annotation tracks automatically triggers extraction and creates indices of sequence databases for all identified biological genomic features (genome sequence, genes, mRNA, ncRNA, CDS and proteins). This offers the possibility to search by keyword or accession number or from a list of identifiers. The results are presented with direct links to the genomic context. In addition, the user can search for similar sequences using an integrated BLAST server. Like for keyword searches, the similarity results show sequence alignments cross-linked to the target genomic regions directly on the genome browser. Finally, myGenomeBrowser provides a form to extract a set of sequences of different genomic features using their identifiers or their genomic coordinates and using the same syntax as JBrowse.

2.4 Managing and sharing data

With myGenomeBrowser, users can share genome browsers in read-only mode by simply filling out a web form with the email addresses of their colleagues. The access control combines an e-mail address, a key and a password. A JSON file containing emails and keys manages the 'authorization' layer. An 'htpasswd' file containing emails and passwords manages the 'authentication' layer. Recipients receive a link (encoding the key) and a password that gives them personalized access to the genome browser and associated mining tools. The owner of the original data may at any time revoke access by removing the corresponding email address from the same web form. In addition, the owner may at any time remove tracks or entire genome browsers and all associated indices via a management form.

3 Conclusion

myGenomeBrowser does not replace institutional genome browsers [UCSC, TAIR (Lamesch *et al.*, 2012), Ensembl (Yates *et al.*, 2016), WormBase (Howe *et al.*, 2016)]. myGenomeBrowser is a complementary solution to meet user needs in autonomously viewing, analyzing, querying and sharing results obtained on their organism of interest, via a range of tools. The environment can be used by everyone, but is particularly designed for platforms or laboratories wishing to increase the autonomy of their biologists working on many organisms, while simplifying system administration of multiple genome browsers. myGenomeBrowser does not require important computational resources. A large dataset like the human genome sequence and its annotation can be indexed in less than one hour on a standard workstation (Intel i7-4600U CPU @ 2.10GHz, RAM: 16Go).

Acknowledgements

We would like to thank Adeline Simon, Nicolas Lapalu, Ludovic Legrand, Ludovic Cottret, Maude Marechaux, Erika Sallet, Olivier Filangi, Cyril Dutech, Clare Gough and reviewers for testing and feedback.

Funding

This work was supported by the BBRIC network (INRA/SPE).

Conflict of Interest: none declared.

References

- Buels,R. *et al.* (2016) JBrowse: a dynamic web platform for genome visualization and analysis. *Genome Biol.*, **17**, 1–12.
- Donlin,M.J. (2009) Using the generic genome browser (GBrowse). *Curr. Protoc. Bioinforma.*
- Ficklin,S.P. *et al.* (2011) Tripal: A construction toolkit for online genome databases. *Database*, **2011**, bar044.
- Freese,N.H. *et al.* (2016) Integrated genome browser: visual analytics platform for genomics. *Bioinformatics* **32**, 2089–2095.
- Goecks,J. *et al.* (2010) Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol.*, **11**, R86.
- Goecks,J. *et al.* (2012) NGS analyses by visualization with Trackster. *Nat. Biotechnol.*, **30**, 1036–1039.
- Gomez,J. *et al.* (2013) BioJS: an open source JavaScript framework for biological data visualization. *Bioinformatics*, **29**, 1103–1104.
- Howe,K.L. *et al.* (2016) WormBase 2016: expanding to enable helminth genomic research. *Nucleic Acids Res.*, **44**, D774–D780.
- Jones,P. *et al.* (2014) InterProScan 5: Genome-scale protein function classification. *Bioinformatics*, **30**, 1236–1240.
- Karolchik,D. *et al.* (2011) The UCSC genome browser. *Curr. Protoc. Hum. Genet.*, doi: 10.1002/0471142905.hg1806s71.
- Lamesch,P. *et al.* (2012) The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Res.*, **40**, D1202–D1210.
- Merkel,D. (2014) Docker: lightweight Linux containers for consistent development and deployment. *Linux J.*, **2014**, Article 2.
- Podicheti,R. *et al.* (2009) WebGBrowse—a web server for GBrowse. *Bioinformatics*, **25**, 1550–1551.
- Robinson,J.T. *et al.* (2011) Integrative genomics viewer. *Nat. Biotechnol.*, **29**, 24–26.
- Yates,A. *et al.* (2016) Ensembl 2016. *Nucleic Acids Res.*, **44**, D710–D716.