bioRxiv preprint doi: https://doi.org/10.1101/2020.05.15.098533; this version posted December 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

BioViz *Connect*: Web application linking CyVerse science gateway resources to genomic visualization in the Integrated Genome Browser

Nowlan H Freese[1]* nfreese@uncc.edu

Karthik Raveendran[1†] kraveend@uncc.edu

Chaitanya Kintali[1†] ckintali@uncc.edu

Srishti Tiwari[1] stiwari8@uncc.edu

Pawan Bole[1] pbole@uncc.edu

Ann E Loraine[1]* aloraine@uncc.edu

[1] Department of Bioinformatics and Genomics, University of North Carolina at Charlotte, Kannapolis, NC 28081, USA

* Corresponding authors

[†] Karthik Raveendran and Chaitanya Kintali contributed equally to this work.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.15.098533; this version posted December 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Abstract

Background: Visualization of genomic data is a key step in validating methods and results. Web-based science gateways such as CyVerse provide storage and analysis tools for genomic data but often lack visualization capability. Desktop visualization tools like Integrated Genome Browser (IGB) enable highly interactive data visualization but are difficult to deploy in science gateways. Developing ways for gateways to interoperate with pre-existing external tools like IGB would enhance their value to users.

Results: We developed BioViz *Connect*, a new web application that *connects* CyVerse and IGB using the CyVerse Terrain Application Programming Interface (API). Using BioViz *Connect*, users can (i) stream their CyVerse data to IGB for visualization, (ii) add IGB specific metadata such as genome version and track appearance to CyVerse data, and (iii) run compute-intensive visual analytics functions to create new visualizations for IGB. To demonstrate BioViz *Connect*, we present an example visual analysis of RNA-Seq data from *Arabidopsis thaliana* plants undergoing heat and desiccation stresses. The example shows how researchers can seamlessly analyze and visualize their CyVerse data in IGB. BioViz *Connect* is accessible from https://bioviz.org.

Conclusions: BioViz *Connect* demonstrates a new way to integrate science gateways with desktop applications using APIs.

Keywords: Integrated Genome Browser, CyVerse, Terrain API, Data Visualization, Data Storage, Genomics

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.15.098533; this version posted December 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Background

Science gateways are Web sites that implement user-friendly interfaces to high performance computing and storage systems that otherwise require specialized expertise to access and use [1]. Science gateways assemble and curate discipline-specific tools within a single, easy-to-use interface for the benefit of researchers in that discipline. In the life sciences, gateways typically feature Web-based interfaces to command-line, Unix-based tools deployed to process and analyze experimental datasets too large to handle on a personal computer, such as data from genome-scale sequencing experiments [2-4]. Researchers upload their data to the gateway site and then use the gateway's interface to develop processing and analysis pipelines customized for their data, performing such tasks as aligning sequences onto a reference genome.

A science gateway provides a single point of access for tools needed to process and analyze data from a research project. However, tools featuring their own graphical user interfaces, such as genome browsers, are difficult to deploy on gateway systems. Many widely used software applications in biology need to run on the user's personal computer [5-8], while others are themselves web-based and require data to reside in tool-specific databases or file systems [9, 10]. In both cases, using these external software tools in conjunction with a science gateway requires a way to move data from the gateway into the external tool. One simple way this can be done is to copy data files onto the system where the external tool resides, but this is often impractical with larger-scale experimental datasets. Another solution keeps the data in place but enables portions of the data to flow into external applications via HTTP range requests. Data are stored in sorted, indexed file formats [8, 11] within the gateway and are made accessible to external systems via URLs. External applications retrieve the index, use it to look up the location of the desired data in the remote file, and then request just the region of the remote file required by the user. In this scenario, the gateway functions as simple Web server.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.15.098533; this version posted December 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

In this paper, we introduce BioViz *Connect*, a Web application that builds on this latter approach to add genome visualization capability to an existing science gateway, the CyVerse Discovery Environment, formerly called iPlant [3, 4]. BioViz *Connect* enables users of Integrated Genome Browser, a highly interactive and mature desktop visualization application implemented in Java [6, 8], to visually analyze their CyVerse data without having to download entire files to their local computer. To start, we describe BioViz *Connect* implementation, focusing on its use of the Terrain API, a REST interface to CyVerse science gateway resources. Next, we describe how BioViz *Connect* enables flow of data into the IGB desktop software by activating a REST API endpoint residing in IGB itself. Lastly, we describe an example use case scenario that highlights how coordinate interaction between remote and local resources enables interactive visualization, which in turn helps users discover biologically meaningful features in data while avoiding artifacts.

Implementation

BioViz *Connect* client and server-side design

BioViz *Connect* consists of two parts: a JavaScript-based user interface that runs in a Web browser and a server-side application that manages authentication and communication with Terrain API endpoints. The user interface code on the client-side is implemented using HTML5, CSS, Bootstrap 4.3.1, JavaScript, and jQuery 1.10.2. The server-side code is implemented in python3 using the Django web application framework (Fig. 1). BioViz *Connect* is deployed on Amazon Web Services infrastructure using the Apache Web server software, but its design is platform agnostic and could run on other cloud or local systems as needed. BioViz *Connect* code is open source and available from https://bitbucket.org/lorainelab/bioviz-connect.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.15.098533; this version posted December 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

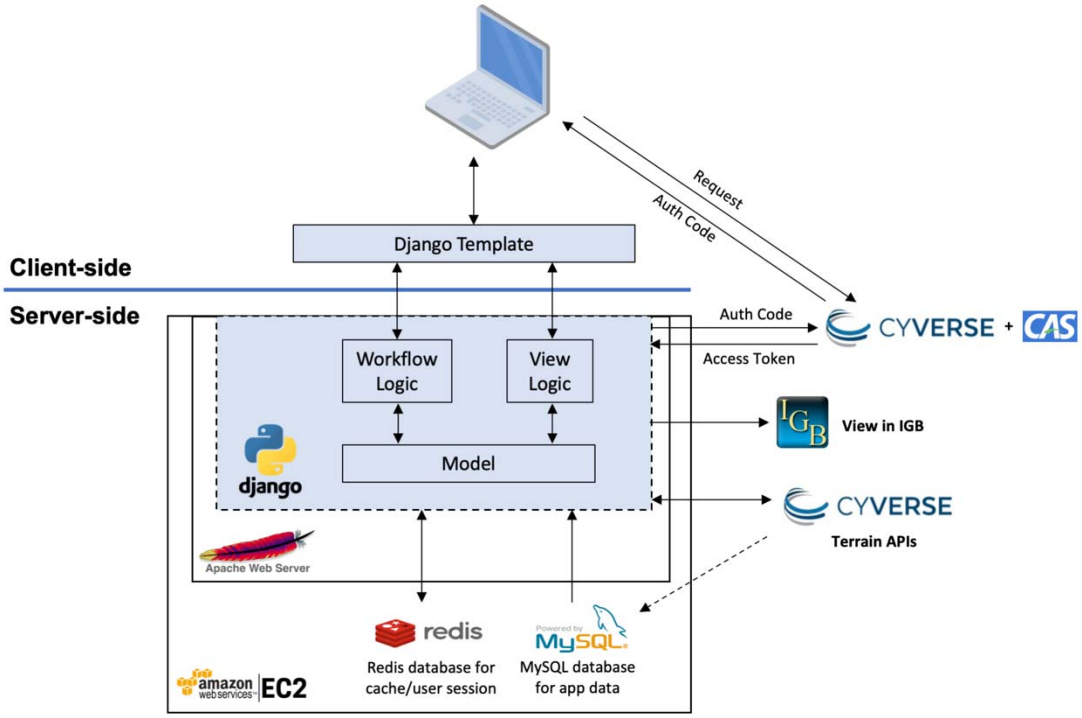


Fig. 1 Diagram illustrating client and server-side design of BioViz *Connect*.

Connect to CyVerse Terrain API

CyVerse is a United States National Science Foundation funded cyberinfrastructure project with the aim of providing computational resources across the life sciences [3, 4]. The CyVerse Discovery Environment is a Web-based science gateway that provides access to data analysis and storage.

A free CyVerse Discovery Environment account is required for accessing the Terrain API for testing and development and can be acquired from https://user.cyverse.org/register. For production, user authentication is handled through Central Authentication Service (CAS) OAuth hosted by CyVerse (Fig. 1). CyVerse infrastructure manages user accounts; no BioViz-specific accounts or passwords are required. Once CAS has authenticated the user's name and password an authorization code is returned as a parameter of the callback URL. BioViz *Connect* uses the authorization code along with the BioViz *Connect* client id and secret to obtain an

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.15.098533; this version posted December 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

access token from Terrain (Fig. 1). The BioViz *Connect* server-side code uses a Redis database to store a user's access token during a session, which is identified by a session id stored in the database and in the user's Web browser. This access token can then be used to authorize calls to the Terrain API on behalf of the user.

Access tokens are valid for 16 hours or until the user chooses to sign out. When a user signs out, BioViz *Connect* deletes the access token from its server-side database, deletes a BioViz *Connect* Web browser cookie with session identifier from the user's Web browser, and then redirects the user back to a CyVerse CAS logout page to notify CyVerse that the access token is no longer valid.

Connect to Integrated Genome Browser

Integrated Genome Browser is a free, open source desktop software program written in Java which users download and install on their local computer systems [6, 8]. Installers for Linux, MacOS, and Windows platforms are available at https://bioviz.org. The IGB source code resides in a git repository hosted on Atlassian's bitbucket.org site (https://bitbucket.org/lorainelab/integrated-genome-browser). IGB version 9.1.4 or greater is required for IGB to connect to BioViz *Connect*.

IGB contains a simple Web server configured to respond to REST-style queries on an IGB-specific port on the user's local computer. Clicking the "View in IGB" button in BioViz *Connect* causes JavaScript code, running in the Web browser, to send a command to IGB running on the desktop. The command includes parameters that tell IGB where to find the data on the CyVerse site along with metadata parameters including genome version and style information. If IGB is not currently running, then the JavaScript code instead displays a notice informing the user that they must first start IGB.

BioViz *Connect* metadata

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.15.098533; this version posted December 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

The Terrain API represents metadata items as three values - Attribute, Value, and Unit. Within BioViz *Connect*, we use a custom, IGB-specific Unit to retrieve and display IGB-relevant metadata for files and folders. We use the attribute to specify what the metadata contains: genome association, foreground color, background color, and comments. The value is appropriate for the attribute type, such as an IGB formatted genome identifier value for the genome association attribute. A list of IGB formatted genome identifiers is populated from the contents.txt metadata file we keep up to date on the IGB quickload site (http://igbquickload.org/quickload/contents.txt). Metadata associated with files are passed to the IGB REST endpoint when a user clicks "View in IGB". The metadata are parsed and then used by IGB to overlay the data onto the correct reference assembly and to specify how the file's data will appear.

Public data visualization

The flow of data from CyVerse into IGB depends on data files having publicly accessible URLs hosted on servers that IGB can use to retrieve data. It also depends on the server software supporting HTTP byte range requests, which allows client software to request arbitrarily small or large sections using byte addresses and thus avoid downloading entire files during the visualization session. Similar to sharing in commercial cloud storage products, users can choose to make individual files public or private through the “Manage Link” right-panel display (Fig. 2). Public links are viewable to anyone with the link, including non-CyVerse users. As a convenience for users, clicking the "View in IGB" button on private files triggers a dialog for creating the public URL for the data set.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.15.098533; this version posted December 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

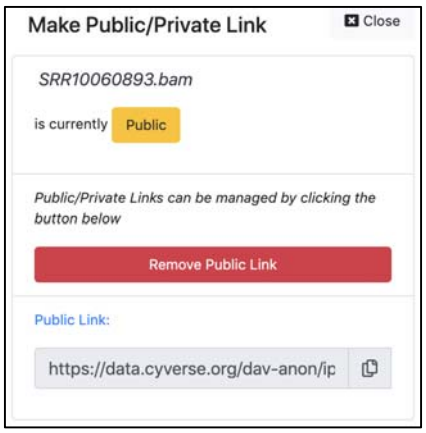


Fig. 2 Right-panel showing Public/Private options for a file. The file is currently set as public, with a public link displayed at the bottom.

BioViz *Connect* data analysis

The CyVerse ecosystem has a large number of Apps for analyzing data. BioViz *Connect* displays a curated subset of published CyVerse Apps that produce output relevant to visualization in IGB. The BioViz *Connect* team controls which apps appear by adding them to the IGB Community within the CyVerse Discovery Environment. An endpoint within BioViz *Connect* retrieves the app descriptions for all apps within the IGB Community and updates the MySQL database (Fig. 1). When a user selects a file and clicks "Analyse", BioViz *Connect* checks the MySQL database to determine which apps can accept the file as input and then displays those apps in the right panel. Once a user has selected an App, BioViz *Connect* dynamically creates an entry form using HTML5 with options for running the App. App jobs are then run on CyVerse cloud computing resources.

RNA-Seq dataset

Example RNA-Seq data are from a study on the effect of heat and desiccation stress on gene expression and RNA splicing in 21-day old Arabidopsis plants. Sequence data are publicly available from the Sequence Read Archive [12] as Bioproject PRJNA509437. Reads were

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.15.098533; this version posted December 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

aligned against the most recent (June 2009) release of the *Arabidopsis thaliana* reference genome, also called "TAIR10" and "TAIR9" using Tophat 2.1.1 [13, 14].

Results

Example use case

To demonstrate BioViz *Connect*, we describe the following use case scenario showing how to annotate, analyze, and visualize an example RNA-Seq data set hosted in the CyVerse Discovery Environment. For the convenience of readers, we published the example data to the CyVerse Discovery Environment "Community" folder, where all CyVerse users can view the files.

Sign in to BioViz *Connect*

The BioViz *Connect* launch page is https://bioviz.org/connect.html. Clicking the "Sign in with your CyVerse ID" link at BioViz *Connect* takes us to the CyVerse hosted Central Authentication Service where we enter our CyVerse username and password. CyVerse DE Accounts are provided free at https://user.cyverse.org/register.

Navigate BioViz *Connect*

Once logged in, our Web browser displays the BioViz *Connect* user interface, which shows a browsable, sortable, paginated view of our account's CyVerse home directory and its contents (Fig. 3A). This view of files and data resembles the interface for modern cloud storage browsers. Clicking a file or folder selects it, and double-clicking a folder opens it and displays the contents. A bread crumb display at the top of the panel shows the path from the root folder to the currently opened folder. The browser's URL bar reflect our current location, and we can use the browser's forward, back, and bookmarking functions to navigate the virtual file system.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.15.098533; this version posted December 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

A search bar at the top of the page lets us search for files or folders by name. Searches use implicit wild cards. For example, we can search for ".bam" to retrieve all binary alignment (bam format) files in the home folder. Only files for which we have read access are returned.

Navigation buttons linking home, shared and community data folders appear in a side panel left of the central file and folder view. The shared folder contains data shared with us by other users. The community folder contains data published for all CyVerse users. The example RNA-Seq data set we are investigating has been published in the community folder. Within the community folder, the path to the data is "BioViz / rnaseq / A_thaliana_Jun_2009 / SRP220157 / reads."

View data in IGB

To illustrate using BioViz *Connect*, we first discuss RNA-Seq alignment files SRR10060893.bam and SRR10060894.bam, named for their corresponding "run" accessions in the Sequence Read Archive. They both represent control samples and are biological replicates of each other. The BioViz *Connect* middle panel view shows that SRR10060893.bam has size 1.61 GB, about twice the size of SRR10060894.bam, which is 0.669 GB. Using their "run" ids, we can look them up in the Sequence Read Archive, where we find that sample SRR10060893 produced around 37.6 million sequences and sample SR1006094 produced 36.7 million sequences. This shows that the two samples were sequenced to about the same depth, and so their resulting alignment files ought to have similar sizes. The fact that they do not suggests there could be a problem with one or both samples. Visualizing the data in Integrated Genome Browser will likely expose any possible problems with the data.

To view and modify BioViz *Connect*-related metadata for a file, we right-click the file name in the middle panel and select "Metadata". This opens a right-side panel with IGB-specific metadata (right side of Fig. 3A). The most important metadata for genomic data files is shown at the top of the right panel display: species and reference genome assembly names. If these are

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.15.098533; this version posted December 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

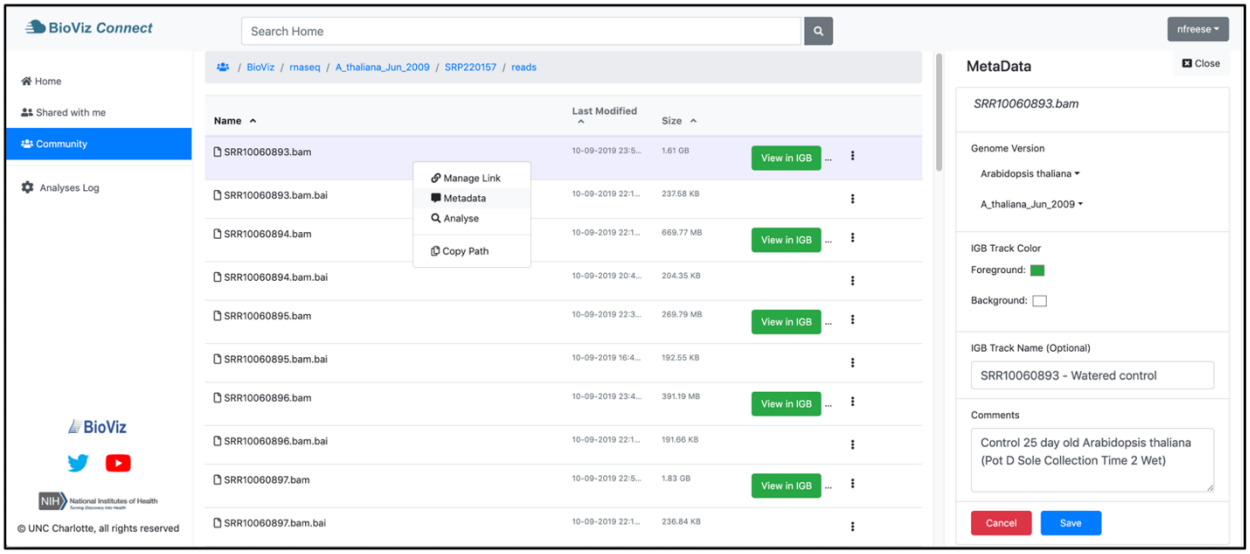
not correctly set, IGB will fail to switch to the correct genome assembly version when we click the "View in IGB" option. Style metadata control how the track will look in IGB, such as track colors and track name. A free-text comment field lets the file's owner describe in human-readable terms what the file represents. If our account has write-permission to the file, we can change the metadata values, thus using the BioViz *Connect* user interface to control how IGB will represent the data once loaded.

To view the data, we start Integrated Genome Browser version 9.1.4 or later. Next, we return to BioViz *Connect* in our Web browser (currently support Chrome, Firefox, or Edge) and click the green button labeled "View in IGB". IGB then opens the correct genome version and adds the file as a new track to its display. Note that the speed in which the track is added will be dependent on internet speeds and can take time to load. We know from prior work that the gene SR45a, encoding an RNA-binding protein, exhibits interesting splicing and expression patterns under stress [15, 16], and so we use IGB's search function to zoom and pan to SR45a (Arabidopsis Genome Initiative code AT1G07350).

To load the alignments into IGB, we click the "Load Data" button at the top right of the IGB window. Once the data load, we immediately notice a problem with SRR10060894 (Fig. 3B). The alignments for this sample appear in vertical stacks covering approximately 30% of the gene's exons. By contrast, the alignments for the other sample fully cover the gene body and include many spliced reads split across introns. The sparser pattern observed in SRR10060894 typically arises when the library synthesis process included too many polymerase chain reaction amplification cycles, which reduces the diversity of resulting sequence data. Unfortunately, this means we must not use this file for further analysis. This example illustrates the importance of visualizing data at multiple steps in the data analysis process.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.15.098533; this version posted December 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

A



B

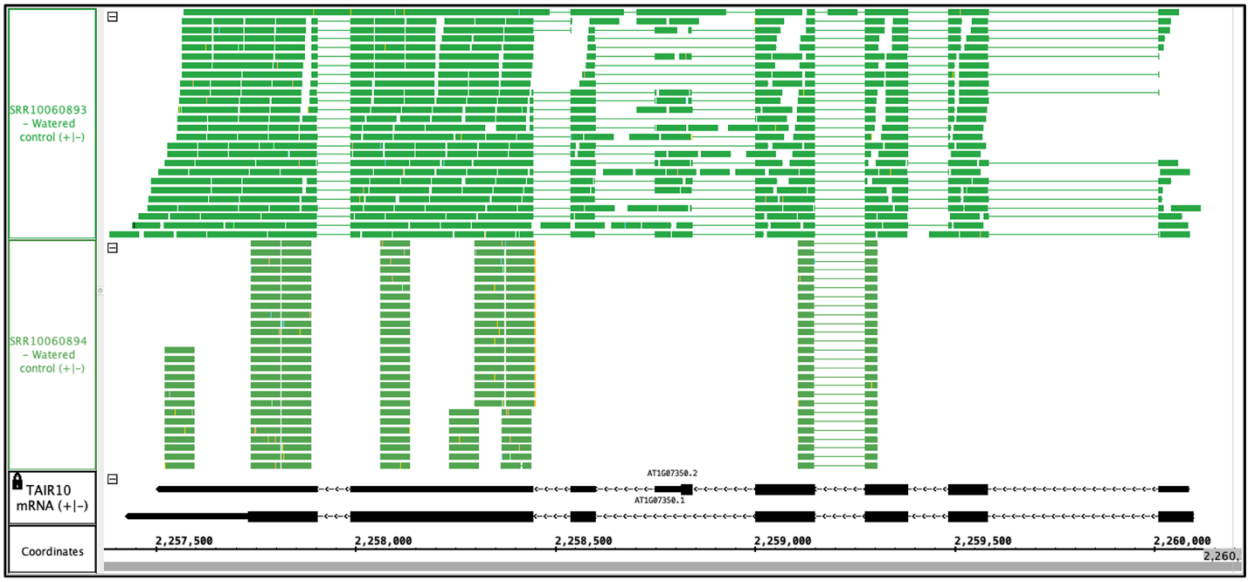


Fig. 3 Example BioViz *Connect* main page and data visualization. (A) BioViz *Connect* main page. The left panel shows shortcuts to home, shared, community folders. The middle panel lists files and folders. The right panel shows the selected file's metadata. (B) SRR10060893.bam and SRR10060894.bam files viewed in IGB overlapping the SR45a gene of *Arabidopsis thaliana*.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.15.098533; this version posted December 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

IGB visual analytics

In addition to the above example, the experiment includes another two files from biological replicates with different sizes. The file SRR10060911.bam is 1.83 Gb, but SRR10060912.bam is only 0.454 Gb. Viewing the "Notes" section in the metadata for each file shows that both samples are from plants that underwent a 3-hour, heat stress treatment. When we open and view these samples in IGB, we see that the pattern of alignments within them is similar, but there are fewer alignments from the smaller file (Fig. 4).

To summarize the amount of sequencing that was done, we use visual analytics within IGB to create coverage graphs using data from the read alignment tracks. To make a coverage graph, we right-click a track label for a read alignment track and choose option "Track Operations > Depth Graph (All)". This generates a new track showing a graph in which the y-axis indicates the number of sequences aligned per position indicated on the x-axis, corresponding to base pair positions in the genome. After modifying the y-axis lower and upper boundary values (using controls in IGB's Graph tab), we see that although the pattern of alignments was similar between the two samples, the overall level of sequencing was different. Sample SRR10060912 has less read coverage across SR45a than SRR10060911 (Fig. 4). Thus, the file size difference most likely reflects different amounts of sequencing rather than a problem with the library synthesis, as was the case in the previous example.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.15.098533; this version posted December 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

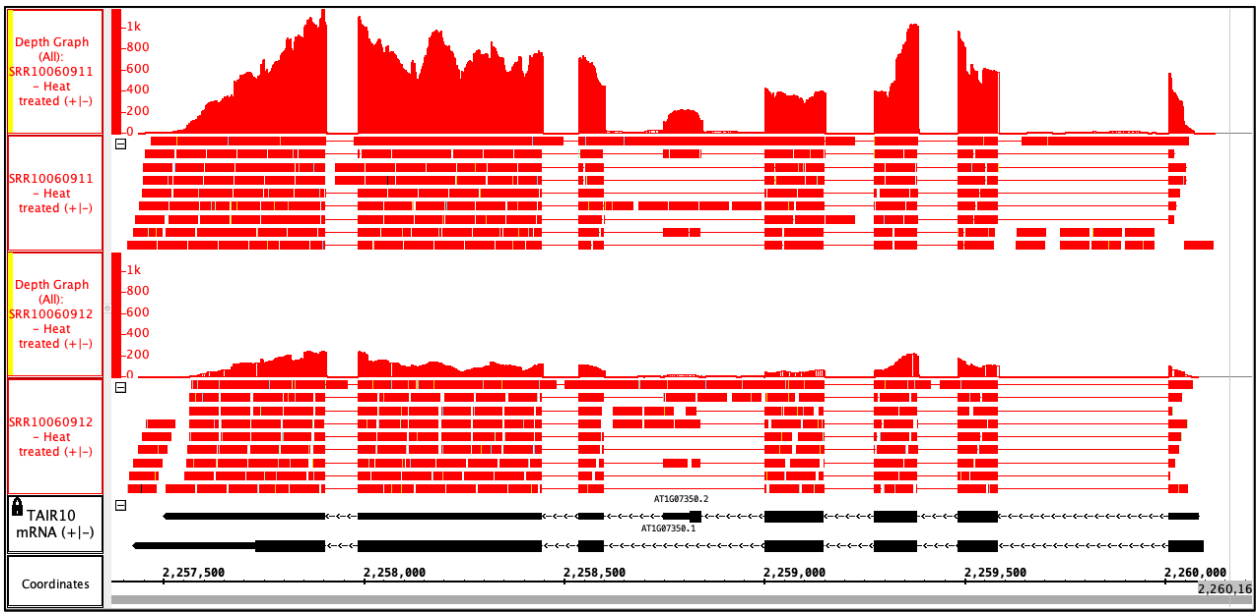


Fig. 4 Heat treated samples viewed in IGB. Alignment tracks and their derived depth graph tracks are shown. The y-axis values represent the number of aligned sequences per base pair position indicated on the coordinates track. Gene models are from the SR45a gene.

Analyze data in BioViz *Connect*

Summarizing read alignments with coverage graphs can provide a fast way to compare expression across sample types, but only if the libraries were sequenced to approximately the same depth. To properly use coverage graphs to assess expression, we need to scale or normalize the graphs to account for differences in sequencing depth. Performing this transformation within IGB itself is impractical, however, as it would require downloading, reading, and processing the entire bam-format alignments file. A much better approach is to off-load computationally intensive visual analytics tasks to CyVerse cloud computing resources. To do this, BioViz *Connect* uses CyVerse apps to generate scaled genome-wide coverage graphs for researchers without taxing their local resources.

To create a scaled coverage graph, we return to BioViz *Connect*, right-click a bam format file, and choose "Analyse." This opens the Analysis right-panel display, which lists

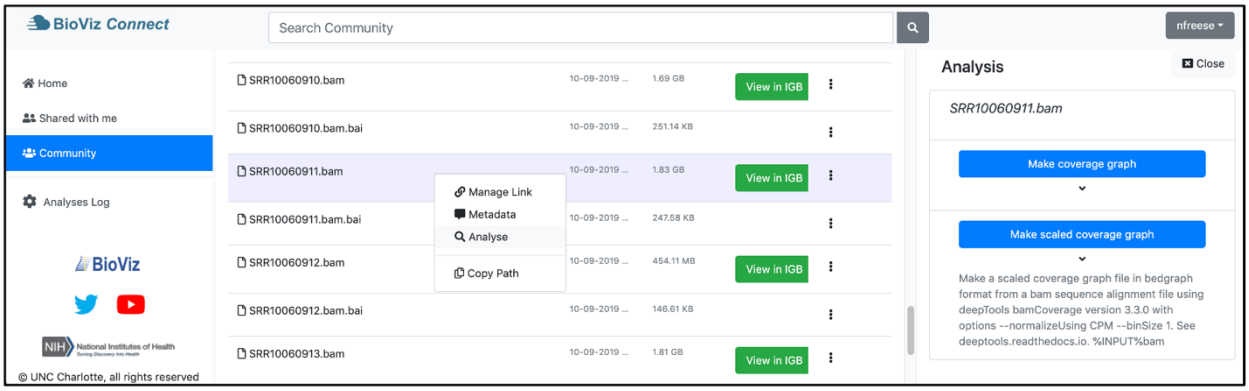
bioRxiv preprint doi: https://doi.org/10.1101/2020.05.15.098533; this version posted December 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

CyVerse Apps that can accept the selected file type as input (Fig. 5A). Selecting "Make scaled coverage graph" opens a form with options for creating the graph using the genomeCoverage function from the DeepTools suite of python-based command line genome analysis tools (Fig. 5B) [17]. We have the option to enter a name for our analysis and the output file. The input file has been automatically filled out, but we need to enter the path to the BAM file's companion index (bai) file. We can then click Run Analysis, which submits a request to run the App using the CyVerse cloud computing resources. We can check the status of our analysis by looking it up in the Analyses Log (Fig. 5C). Analyses are listed as Queued, Running, Failed, or Completed. The CyVerse infrastructure sends an email once the analysis has finished. The output will be located in the analyses folder in our home directory, as we do not have write permission to the Community folder. If the input file were located in the home folder, the output would be located in the same location as the input file. Clicking the analysis name in the Analyses Log opens a shortcut to the folder where the output data are stored.

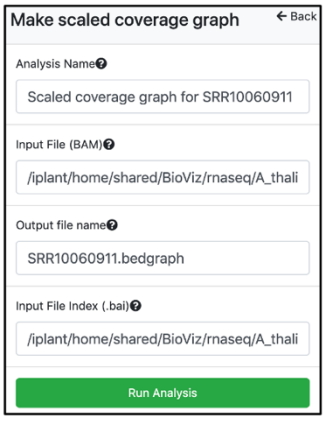
Figure 5D shows scaled coverage graphs created for the two RNA-Seq samples described above. Examining the scaled coverage graphs for SR45a shows that both heat-treated samples had similar scaled expression (Fig. 5D). This indicates that the difference observed in the unscaled coverage graph was likely an artifact of lower sequencing depth rather than depressed expression of the gene in one replicate versus the other. Figure 5D also shows a scaled coverage graph (labeled "cool") from a control sample that received no heat stress treatment. The control sample's scaled coverage graph exhibits approximately 10-fold lower expression (measured as y-axis values) across the spliced regions of the gene, visually demonstrating that the heat stress treatment increased SR45a transcript abundance during stress [15].

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.15.098533; this version posted December 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

A



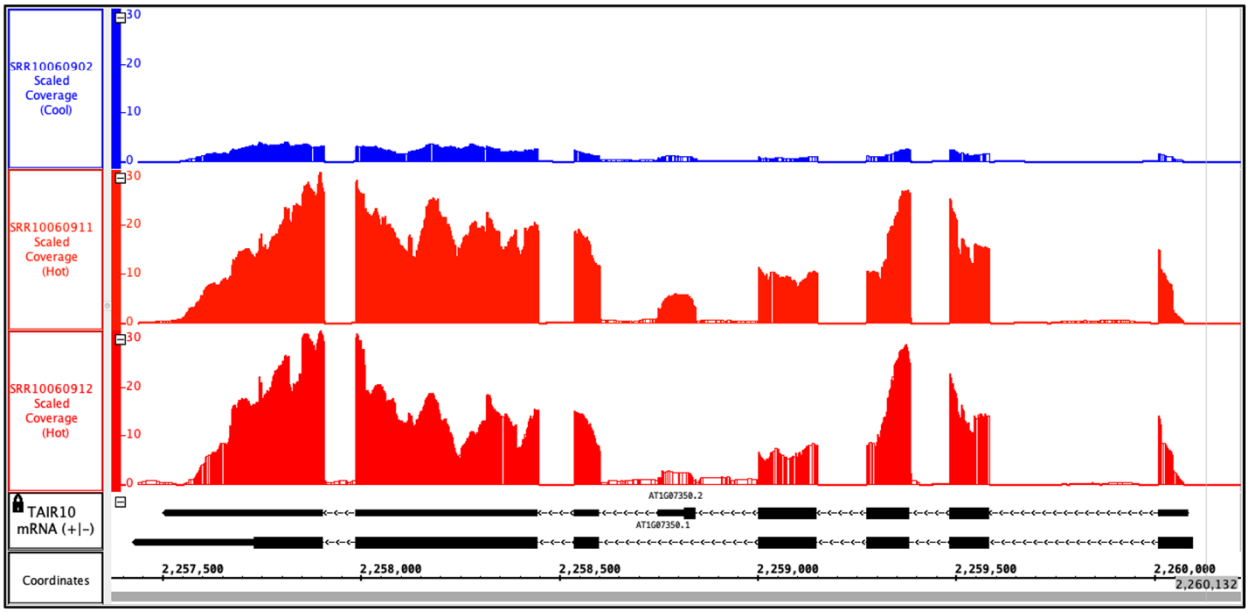
B



C

Name	App Name	Start Date	End Date	Status
Scaled coverage graph for SRR10...	Make scaled coverage graph	05-01-2020 19:07:57	null	Submitted
Scaled coverage graph for SRR10...	Make scaled coverage graph	05-01-2020 19:07:11	null	Running
Scaled Coverage Graph for SRR1...	Create Scaled Coverage Graph	04-28-2020 19:47:26	04-28-2020 19:58:40	Completed
Scaled Coverage Graph for SRR1...	Create Scaled Coverage Graph	04-28-2020 19:46:28	04-28-2020 19:58:40	Completed
Not public read access		10-16-2019 20:55:51	10-16-2019 21:13:28	Completed
Not Public Read Access 3	FastQC 0.11.5 (multi-file)	10-16-2019 20:58:34	10-16-2019 21:00:19	Completed
Not Public Read Access 2	FastQC 0.11.5 (multi-file)	10-16-2019 20:56:20	10-16-2019 20:57:39	Failed
PK_Unzip_analysis1	PK Unzip	02-25-2015 18:53:57	02-25-2015 19:03:48	Completed
Uncompress_with_bunzip2_analys...	Uncompress with bunzip2	02-25-2015 18:50:52	02-25-2015 18:52:15	Failed
Uncompress_files_with_gunzip_an...	Uncompress files with gunzip	02-25-2015 18:47:24	02-25-2015 18:48:24	Failed
Cuffmerge2_analysis1	Cuffmerge2	01-30-2015 14:33:32	01-30-2015 14:40:08	Failed
Cufflinks2_analysis1	Cufflinks2	01-29-2015 20:28:54	01-29-2015 21:23:53	Completed
TopHat2-SE_analysis1	TopHat2-SE	01-29-2015 18:14:44	01-29-2015 19:23:06	Failed

D



bioRxiv preprint doi: https://doi.org/10.1101/2020.05.15.098533; this version posted December 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Fig. 5 Example analysis in BioViz *Connect* with output visualized in IGB. (A) BioViz *Connect* main page with analysis right panel open. (B) Scaled coverage graph analysis options for naming the analysis, selecting input file, output file name, and index file selection. (C) Analyses log showing the status of current and previous jobs. (D) SRR10060902 (control), SRR10060911 (heat treated), and SRR10060912 (heat treated) scaled bedgraph files viewed in IGB overlapping the SR45a gene of *Arabidopsis thaliana*.

Conclusions

BioViz *Connect* introduces and demonstrates several innovations in the field of science gateway development and research. First, BioViz *Connect* enables users of the CyVerse science gateway to visualize genomic data files from their CyVerse accounts using Integrated Genome Browser, a desktop application. To our knowledge, BioViz *Connect* is the first and only resource that enables CyVerse users to visualize data in a genome browser or other desktop tool. Second, we implemented BioViz *Connect* using the CyVerse Terrain API, a collection of REST endpoints that form a comprehensive computational interface to CyVerse resources. As such, BioViz *Connect* is the first application developed using the Terrain API by a group outside the CyVerse development team and using an entirely different visual interface. As such, BioViz *Connect* demonstrates to the larger community of biologists, developers, and funders that modern, feature-rich REST interfaces to powerful computational resources stimulate and enable innovation and progress. Third, BioViz *Connect* introduces a new breed of science gateway specialized for visualization on the desktop. Rather than re-implement visualization tooling within the gateway, we instead created a new kind of gateway that manages and coordinates the flow of data into an-already mature and robust desktop application with a pre-existing userbase. Fourth, BioViz Connect demonstrates using powerful remote resources to perform compute-intensive visual analytics functions, as in the genome-wide coverage graph example

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.15.098533; this version posted December 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

described in preceding sections of the paper. This example is straightforward and uses a pre-existing algorithm, but it shows how developers can harness a more powerful gateway system to develop and deploy all-new visualizations of genomic data.

Availability and requirements

Project name: BioViz *Connect*

Project home page: https://bitbucket.org/lorainelab/bioviz-connect

Operating system(s): Linux-based systems for the BioViz *Connect* server. BioViz Connect web app tested on Ubuntu, Windows 10, MacOS Mojave operating systems and Chrome, Firefox, and Edge web browsers.

Programming language: HTML, JavaScript, CSS, and python3

Other requirements: A free CyVerse account is required to login: https://user.cyverse.org/register. Integrated Genome Browser 9.1.4 or greater required for visualizing data.

License: Common Public License Version 1.0

Any restrictions to use by non-academics: No

Abbreviations

API: Application Programming Interface; CSS: Cascading Style Sheets; HTML: HyperText Markup Language; IGB: Integrated Genome Browser; REST: REpresentational State Transfer

Declarations

Availability of data and Materials

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.15.098533; this version posted December 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

RNA-Seq data files are available from https://data.cyverse.org/dav-anon/iplant/home/shared/BioViz/rnaseq.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing Interests

The authors declare that they have no competing interests.

Funding

Research reported in this publication was supported by the National Institute of General Medical Sciences of the National Institutes of Health under award number 5R01GM121927-02. Funding was used to plan, design, and develop the software reported in the article.

Authors' Contributions

NHF and AEL conceived of and supervised the project. KR, CK, ST, and PB planned and developed BioViz *Connect*. NHF, AEL, KR, CK, ST, PB tested and debugged BioViz *Connect*. NHF, KR, CK, and AEL wrote the draft manuscript. All authors read and approved the final manuscript.

Acknowledgements

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.15.098533; this version posted December 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

We thank Paul Sarando, Sarah Roberts, Sriram Srinivasan, Ian McEwen, Ramona Walls, and Reetu Tuteja for their assistance with the Terrain API and publishing CyVerse apps, which was made possible through CyVerse's External Collaborative Partnership program.

References

1. Wilkins-Diehr N, Gannon D, Klimeck G, Oster S, Pamidighantam S: TeraGrid Science Gateways and Their Impact on Science. *Computer* 2008, 41(11):32-41.
2. Giardine B, Riemer C, Hardison RC, Burhans R, Elnitski L, Shah P, Zhang Y, Blankenberg D, Albert I, Taylor J *et al*: Galaxy: a platform for interactive large-scale genome analysis. *Genome research* 2005, 15(10):1451-1455.
3. Goff SA, Vaughn M, McKay S, Lyons E, Stapleton AE, Gessler D, Matasci N, Wang L, Hanlon M, Lenards A *et al*: The iPlant Collaborative: Cyberinfrastructure for Plant Biology. *Front Plant Sci* 2011, 2:34-34.
4. Merchant N, Lyons E, Goff S, Vaughn M, Ware D, Micklos D, Antin P: The iPlant Collaborative: Cyberinfrastructure for Enabling Data to Discovery for the Life Sciences. *PLoS Biol* 2016, 14(1):e1002342-e1002342.
5. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T: Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research* 2003, 13(11):2498-2504.
6. Nicol JW, Helt GA, Blanchard SG, Jr., Raja A, Loraine AE: The Integrated Genome Browser: free software for distribution and exploration of genome-scale datasets. *Bioinformatics* 2009, 25(20):2730-2731.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.15.098533; this version posted December 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

7. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP: Integrative genomics viewer. *Nature Biotechnology* 2011, 29(1):24-26.

8. Freese NH, Norris DC, Loraine AE: Integrated genome browser: visual analytics platform for genomics. *Bioinformatics* 2016, 32(14):2089-2095.

9. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler, David: The Human Genome Browser at UCSC. *Genome research* 2002, 12(6):996-1006.

10. Buels R, Yao E, Diesh CM, Hayes RD, Munoz-Torres M, Helt G, Goodstein DM, Elsik CG, Lewis SE, Stein L *et al*: JBrowse: a dynamic web platform for genome visualization and analysis. *Genome biology* 2016, 17(1):66.

11. Blankenberg D, Chilton J, Coraor N: Galaxy External Display Applications: closing a dataflow interoperability loop. *Nature methods* 2020, 17(2):123-124.

12. Leinonen R, Sugawara H, Shumway M: The sequence read archive. *Nucleic Acids Res* 2011, 39(Database issue):D19-21.

13. Berardini TZ, Reiser L, Li D, Mezheritsky Y, Muller R, Strait E, Huala E: The Arabidopsis information resource: Making and mining the "gold standard" annotated reference plant genome. *Genesis (New York, NY : 2000)* 2015, 53(8):474-485.

14. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL: TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome biology* 2013, 14(4):R36.

15. Gulledge AA, Roberts AD, Vora H, Patel K, Loraine AE: Mining Arabidopsis thaliana RNA-seq data with Integrated Genome Browser reveals stress-induced alternative splicing of the putative splicing regulator SR45a. *American journal of botany* 2012, 99(2):219-231.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.15.098533; this version posted December 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

16. Yoshimura K, Mori T, Yokoyama K, Koike Y, Tanabe N, Sato N, Takahashi H, Maruta T, Shigeoka S: Identification of alternative splicing events regulated by an Arabidopsis serine/arginine-like protein, atSR45a, in response to high-light stress using a tiling array. *Plant & cell physiology* 2011, 52(10):1786-1805.

17. Ramirez F, Ryan DP, Gruning B, Bhardwaj V, Kilpert F, Richter AS, Heyne S, Dundar F, Manke T: deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res* 2016, 44(W1):W160-165.