bioRxiv preprint doi: https://doi.org/10.1101/2020.10.09.327114; this version posted October 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Authors

Nagampalli VijayKrishna[1], Jayadev Joshi[1], Nate Coraor[2], Jennifer Hillman-Jackson[2], Dave Bouvier[2], Marius van den Beek[2], Ignacio Eguinoa[3,4], Frederik Coppens[3,4], Sergey Golitsynskiy[5], Michał Stolarczyk[6], Nathan C. Sheffield[6], Simon Gladman[7], Gianmauro Cuccuru[8], Björn Grüning[8], Nicola Soranzo[9], Helena Rasche[10], Bradley W. Langhorst[11], Matthias Bernt[12], Dan Fornika[13], David Anderson de Lima Morais[14], Michel Barrette[14], Peter van Heusden[15], Mauro Petrillo[16], Antonio Puertas-Gallardo[16], Alex Patak[16], Hans-Rudolf Hotz[17,18], and Daniel Blankenberg[1,19,*]

[1]Genomic Medicine Institute, Cleveland Clinic, Cleveland, Ohio, USA

[2]Department of Biochemistry and Molecular Biology, Penn State University, University Park, PA, USA

[3]VIB Center for Plant Systems Biology, Ghent, Belgium

[4]Department of Plant Biotechnology and Bioinformatics, Ghent University, Ghent, Belgium

[5]Department of Biology, Johns Hopkins University, Baltimore, MD, USA

[6]Center for Public Health Genomics, University of Virginia, Charlottesville, VA, USA

[7]University of Melbourne, Melbourne, Australia

[8]University of Freiburg, Freiburg im Breisgau, Germany

[9]Earlham Institute, Norwich Research Park, Norwich, UK

[10]Erasmus Medical Center, Clinical Bioinformatics group, Department of Pathology, Wytemaweg 80, 3015 CN, Rotterdam, The Netherlands

[11]New England Biolabs, MA, USA

[12]Helmholtz Centre for Environmental Research, Young Investigators Group Bioinformatics and Transcriptomics, UFZ Leipzig, Germany

[13]BC Centre for Disease Control Public Health Laboratory, Vancouver, Canada

[14]Centre de Calcul Scientifique, Université de Sherbrooke, Sherbrooke, Canada

[15]South African National Bioinformatics Institute, University of the Western Cape, Cape Town, South Africa

[16]European Commission, Joint Research Centre (JRC), Ispra, Italy

[17]Friedrich Miescher Institute for Biomedical Research, Basel, Switzerland

[18]SIB Swiss Institute of Bioinformatics, Basel, Switzerland

[19]Department of Molecular Medicine, Cleveland Clinic Lerner College of Medicine, Case Western Reserve University, Cleveland, Ohio, USA

*e-mail: blanked2@ccf.org

Title

Expanding the Galaxy's reference data

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.09.327114; this version posted October 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Abstract

Summary: Properly and effectively managing reference datasets is an important task for many bioinformatics analyses. Refgenie is a reference asset management system that allows to easily organize, retrieve, and share such datasets. Here, we describe the integration of refgenie into the Galaxy platform. Server administrators are able to configure Galaxy to make use of reference datasets made available on a refgenie instance. Additionally, a Galaxy Data Manager tool has been developed to provide a graphical interface to refgenie's remote reference retrieval functionality. A large collection of reference datasets has also been made available using the CVMFS repository from GalaxyProject.org, with mirrors across the United States, Canada, Europe, and Australia, enabling easy use outside of Galaxy.

Availability and implementation: The ability of Galaxy to use refgenie assets was added to the core Galaxy framework in version 20.05, which is available from https://github.com/galaxyproject/galaxy under the Academic Free License version 3.0. The refgenie Data Manager tool can be installed via the Galaxy ToolShed, with source code managed at https://github.com/galaxyproject/tools-iuc/data_managers/data_manager_refgenie_pull and released using an MIT license.

Introduction

Among the primary resources required when performing biomedical genomic analyses are the collection of reference datasets and annotations, including sequences and features (such as genes, locations of epigenetic modifications, regulatory regions, etc.) as well as derived datasets such as index files for sequencing-read mapping tools e.g. Bowtie 2[1] or BWA[2], FASTA sequence indexes with SAMtools[3], etc. While generating the derived data is relatively straightforward, such as running "bwa index -a bwtsw reference.fa" or "samtools faidx reference.fa", these important reference data have to be managed efficiently. Moreover, in order to allow for reproducible analyses, not only do genome build versions (e.g. hg19 vs. hg38) tool versions (e.g. indexed with BWA version 0.7.17-r1188) need to be recorded, also algorithmic choices (e.g. algorithm for constructing BWT index: "bwtsw" or "is") are essential. This has led to the development of several software solutions for reference genome resource management.

We have previously developed the Galaxy Data Manager framework[4] which enables provenance-backed and reproducible reference datasets inside of Galaxy[5,6]. The Data Manager framework leverages the Galaxy tool system to enable community-driven creation and dissemination of reference dataset build recipes as Galaxy tools. These Data Manager tools can be installed on-demand from the ToolShed[7] to allow Galaxy administrators to fetch, build, and install new reference data. All underlying tool dependencies (SAMtools, BWA, STAR[8], Kraken[9], etc.) are well defined with versions and can be automatically and reproducibly resolved using e.g. Bioconda[10] and Docker. As with any Galaxy tool, multiple versions of a Data Manager can be installed and executed, with provenance and reproducibility maintained. Data Manager tools can be accessed using the Galaxy GUI, a REST-ful API, or by using command-line scripts. Currently, over 70 Data Manager tools have been created and shared on the ToolShed by the Galaxy community.

Refgenie[11] has been recently published as a reference genome resource manager. Refgenie provides a command-line interface to download, build, and access reference genome assembly

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.09.327114; this version posted October 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

assets (i.e. reference sequences, mapping indexes, etc.). It can also be controlled via an API from other software. There are currently 24 recipes available for building assets locally. Refgenie can also provide a lightweight web server enabling sharing of local resources.

In light of its potential to become a widely accepted standard for reference data management, we have enabled the use of refgenie-based assets inside of Galaxy. Users are able to seamlessly access refgenie-defined references within Galaxy tools, and administrators are able to access refgenie's ability to download reference datasets from remote resources. This approach enables the simultaneous creation and use of reference files by both Data Managers and refgenie within Galaxy.

Moreover, the adoption of refgenie by the Galaxy community will make high quality reference data sets and derived assets available for the broad scientific community.

Methods

We have extended the Galaxy Data Manager framework to interoperate with refgenie, a standalone reference genome manager.

Using refgenie assets in Galaxy

Integration of refgenie with Galaxy is available as of Galaxy version 20.05. Enabling refgenie usage within Galaxy is accomplished by following an optional two-step process: installing refgenie and configuring Galaxy.

A server administrator must first install refgenie using the standard prescribed instructions and take note of the path selected for its genome configuration file (see also https://galaxyproject.org/admin/refgenie/ and supplementary materials).

Second, refgenie is enabled in Galaxy by setting the 'refgenie_config_file' value to the previously chosen genome configuration file path within the primary Galaxy configuration file (e.g. 'galaxy.yml'; figure 1A). This setting enables the default mappings between refgenie asset names and values into Galaxy data table entries. Currently, default mappings have been defined for genome build identifiers, chromosome sizes and sequences, samtools FASTA indexes, Bowtie 2 indexes, BWA indexes, and HISAT2 indexes. Additional mappings can be defined as shown in figure 1B. The assets loaded from refgenie, are merged with those from Galaxy's native Data Managers. This approach provides maximum flexibility for the administrator while creating a single unified access point from the Galaxy user's perspective. In fact, it is possible for an administrator to define multiple different refgenie configurations within a single Galaxy instance, making the assets defined in each resource available as a combined reference collection in Galaxy. This can be helpful, for example, if the administrator wants to provide access to the prebuilt collection of datasets in CVMFS (see below) along with additional locally maintained refgenie datasets.

Once the refgenie integration has been configured, the users of the Galaxy server will be able to select the refgenie-provided datasets in Galaxy tools that make use of the data tables for which a mapping is defined (figure 1C).

Retrieving additional refgenie assets in Galaxy

An initial refgenie installation contains no data. However, the administrator is able to search, download, and install assets into refgenie from public remote servers. Any asset loaded into

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.09.327114; this version posted October 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

refgenie using the standard approaches will be made available within Galaxy. To facilitate management of the assets, a Galaxy Data Manager tool has been created to provide a graphical interface to refgenie, which is now available in the Galaxy ToolShed[7]. When a Galaxy administrator accesses this Data Manager, a list of available remote assets is dynamically populated and displayed (figure 1D). These available references are organized by remote resource, genome build, asset type, and tag. The administrator selects one or more assets to retrieve, and then clicks "Execute". An external process is queued, using Galaxy's job management system, which calls refgenie's standard command-line tools to perform the requested actions. Once refgenie has completed, Galaxy loads the new values and makes them available to users. This Data Manager can also be controlled using the standard Galaxy tool API.

A CVMFS mirror

The refgenie software provides reference asset downloads from http://refgenomes.databio.org/. To increase reliability and availability, we have created a mirror of all of Galaxy and refgenie reference data available in a CernVM-FS[12] (CVMFS) repository. CVMFS is a caching, HTTP-based filesystem with a Filesystem in Userspace (FUSE) (mount) client. The Galaxy CVMFS repository distributes the data in multiple replicas across the United States, Canada, Europe, and Australia. Moreover, it is possible to make the mirrored data directly available on any other server (see https://galaxyproject.org/admin/reference-data-repo/). When initially mounted, CVMFS does not consume any additional local disk space. Instead, as files are accessed, they are pulled from one of the replica (Stratum 1) servers to a local disk-based cache of a configurable size. This allows the administrator to directly load the refgenie configuration from e.g. /cvmfs/refgenomes-databio.galaxyproject.org/genomes_config.yaml, without needing to predetermine which assets the users may desire.

The presented approach of data distribution allows access to the same reference data through Galaxy as well as e.g. command line, enabling reproducible analysis beyond the Galaxy community. Additional optimizations, e.g. using HTTP proxies or alien cache, are easily deployable across data centers, HPC clusters, and local labs, greatly facilitating scalability.

Conclusions

We have extended the Galaxy platform to support the use of externally managed reference datasets using refgenie. Galaxy is able to seamlessly merge refgenie-provided references with those specified within its native Data Manager framework. We have provided a Data Manager tool that enables Galaxy administrators to view and pull additional refgenie assets graphically or using an API. This allows administrators to easily populate a Galaxy instance with standardized references to built-in data, improving interoperability of analysis among the now widely extended network of public Galaxy servers (https://galaxyproject.org/use/). Galaxy administrators are able to use standard Data Managers to build additional reference artifacts based upon refgenie provided asset sources. While these reference datasets will be available to Galaxy users, there is not yet a way to push them back into refgenie. Furthermore, although refgenie contains its own system for building assets, due to the lack of several important features, such as a system for declaring dependency requirements, adding this functionality to the Data Manager framework is left as a future endeavor.

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.09.327114; this version posted October 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

We are pleased to report that this integration between refgenie and Galaxy was achieved without the need for architectural modifications to either tool and hope to see other examples of such collaboration in the future. The availability of these reference datasets via CVMFS, refgenie, Galaxy Data Managers, and other standard web protocols provides a globally-accessible network of provenance-backed reference datasets that are useful, accessible, and interoperable to a diverse range of researchers.

Acknowledgements

The authors would like to thank the communities and contributors of refgenie and Galaxy for their support and for creating the environments that allowed us to build this collaboration. We would also like to thank all who support open source software, and those who generate and distribute datasets following the FAIR principles.

Contributions

All authors contributed, read, and approved the manuscript.

Competing Interest

Daniel Blankenberg and Nate Coraor have a significant financial interest in GalaxyWorks, a company that may have a commercial interest in the results of this research and technology. This potential conflict of interest has been reviewed and is managed by their respective institutions.

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.09.327114; this version posted October 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Figures

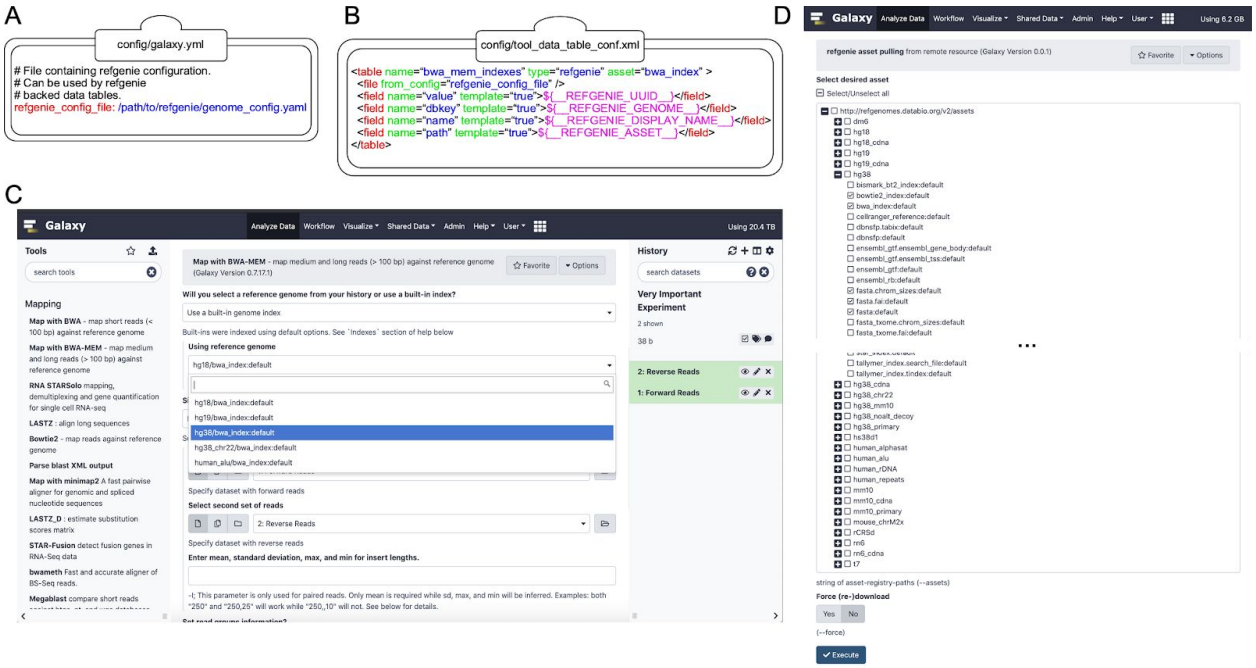

Figure 1. Extending Galaxy's reference data with refgenie. (A) Setting the value of 'refgenie_config_file' to the previously chosen genome configuration file path within the primary Galaxy configuration file (e.g. 'galaxy.yml'). (B) Example data table mapping between refgenie assets and Galaxy data tables for the BWA tool. Cheetah templating language is used to specify mappings between values, with several pre-populated refgenie variables available as shown. (C) refgenie assets are available for users to select and use in the Galaxy BWA tool. In this example, the user is mapping a set of paired-end sequencing reads against the hg38 genome. (D) A dynamically generated list of available remote refgenie assets are listed for an administrator to select in the 'refgenie pull' Galaxy Data Manager tool.

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.09.327114; this version posted October 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

References

1 Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods* 2012;**9**:357–9. https://doi.org/10.1038/nmeth.1923.

2 Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 2009;**25**:1754–60. https://doi.org/10.1093/bioinformatics/btp324.

3 Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 2009;**25**:2078–9. https://doi.org/10.1093/bioinformatics/btp352.

4 Blankenberg D, Johnson JE, Galaxy Team, Taylor J, Nekrutenko A. Wrangling Galaxy's reference data. *Bioinformatics* 2014;**30**:1917–9. https://doi.org/10.1093/bioinformatics/btu119.

5 Giardine B, Riemer C, Hardison RC, Burhans R, Elnitski L, Shah P, *et al.* Galaxy: a platform for interactive large-scale genome analysis. *Genome Res* 2005;**15**:1451–5. https://doi.org/10.1101/gr.4086505.

6 Jalili V, Afgan E, Gu Q, Clements D, Blankenberg D, Goecks J, *et al.* The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2020 update. *Nucleic Acids Res* 2020;**48**:8205–7. https://doi.org/10.1093/nar/gkaa554.

7 Blankenberg D, Von Kuster G, Bouvier E, Baker D, Afgan E, Stoler N, *et al.* Dissemination of scientific software with Galaxy ToolShed. *Genome Biol* 2014;**15**:403. https://doi.org/10.1186/gb4161.

8 Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 2013;**29**:15–21. https://doi.org/10.1093/bioinformatics/bts635.

9 Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol* 2014;**15**:R46. https://doi.org/10.1186/gb-2014-15-3-r46.

10 Grüning B, Dale R, Sjödin A, Chapman BA, Rowe J, Tomkins-Tinch CH, *et al.* Bioconda: sustainable and comprehensive software distribution for the life sciences. *Nat Methods* 2018;**15**:475–6. https://doi.org/10.1038/s41592-018-0046-7.

11 Stolarczyk M, Reuter VP, Smith JP, Magee NE, Sheffield NC. Refgenie: a reference genome resource manager. *Gigascience* 2020;**9.**: https://doi.org/10.1093/gigascience/giz149.

12 Blomer J, Aguado-Sánchez C, Buncic P, Harutyunyan A. Distributing LHC application software and conditions databases using the CernVM file system. *J Phys Conf Ser* 2011;**331**:042003. https://doi.org/10.1088/1742-6596/331/4/042003.