Bioinformatics, 36(15), 2020, 4357–4359
doi: 10.1093/bioinformatics/btaa556
Advance Access Publication Date: 3 June 2020
Applications Note

Sequence analysis

GLASSgo in Galaxy: high-throughput, reproducible and easy-to-integrate prediction of sRNA homologs

Richard A. Schäfer[1], Steffen C. Lott[2], Jens Georg[2], Björn A. Grüning[3], Wolfgang R. Hess[2], Björn Voß[1]*

[1]Computational Biology, Institute of Biochemical Engineering, University of Stuttgart, Stuttgart 70569, Germany, [2]Genetics and Experimental Bioinformatics, Institute of Biology III, University of Freiburg, Freiburg 79104, Germany and [3]Bioinformatics, Institute of Computer Science, University of Freiburg, Freiburg 79110, Germany

*To whom correspondence should be addressed.

Associate Editor: Jan Gorodkin

Received on January 21, 2020; revised on May 13, 2020; editorial decision on May 27, 2020; accepted on May 29, 2020

Abstract

Motivation: The correct prediction of bacterial sRNA homologs is a prerequisite for many downstream analyses based on comparative genomics, but it is frequently challenging due to the short length and distinct heterogeneity of such homologs. GLobal Automatic Small RNA Search go (GLASSgo) is an efficient tool for the prediction of sRNA homologs from a single input query. To make the algorithm available to a broader community, we offer a Docker container along with a free-access web service. For non-computer scientists, the web service provides a user-friendly interface. However, capabilities were lacking so far for batch processing, version control and direct interaction with compatible software applications as a workflow management system can provide.

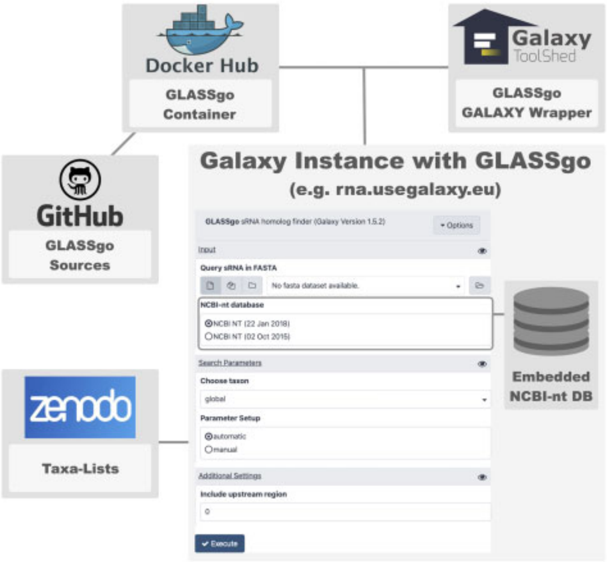
Results: Here, we present GLASSgo 1.5.2, an updated version that is fully incorporated into the workflow management system Galaxy. The improved version contains a new feature for extracting the upstream regions, allowing the search for conserved promoter elements. Additionally, it supports the use of accession numbers instead of the outdated GI numbers, which widens the applicability of the tool.

Availability and implementation: GLASSgo is available at https://github.com/lotts/GLASSgo/ under the MIT license and is accompanied by instruction and application data. Furthermore, it can be installed into any Galaxy instance using the Galaxy ToolShed.

Contact: bjoern.voss@ibvt.uni-stuttgart.de

1 Introduction

One of the most fundamental analyses of newly discovered genes is the search for potential homologues in related species and beyond. Regulatory small RNAs (sRNAs) are often important post-transcriptional regulators of gene expression in bacteria. For their better characterization, information about homologs existing elsewhere is crucially relevant. However, the identification of homologous sRNA genes can be challenging due to their relatively short length, frequently little sequence conservation and the absence of reading frames and signals for translation. Therefore, we developed GLobal Automatic Small RNA Search go (GLASSgo) for the fast and reliable search for sRNA homologs starting from a single sRNA query (Lott *et al.*, 2018). Since its initial release in 2018, this algorithm has filled a gap in the computational analysis of sRNAs in bacteria (Wright et al., 2018). Briefly, GLASSgo performs iterative, sensitive searches for sequence similarity using BLASTn, filtering and clustering, with an optional final structure based assessment. GLASSgo is available on GitHub, Docker Hub (https://hub.docker.com/r/lotts/glassgo) and as a web service (http://rna.informatik.uni-freiburg.de/GLASSgo/) (Raden *et al.*, 2018). Especially the latter two options greatly simplify the use of GLASSgo by non-bioinformaticians. However, results may be difficult to reproduce later because GLASSgo makes heavy use of external resources, such as the NCBI BLAST databases that are updated regularly. Furthermore, high-throughput analyses for hundreds or thousands of sRNAs, and the integration of GLASSgo into automated workflows is not trivial. A framework that provides both, workflow automation and reproducibility, is Galaxy (Afgan *et al.*, 2018). For tool integration, Galaxy offers a one-click installation solution, which automatically resolves all dependencies and keeps track of all versions (libraries, tools and XML wrapper versions). Here, we present an updated version of GLASSgo and its integration into the Galaxy workflow management system utilizing Docker technology.



GLASSgo sRNA homolog finder (Galaxy Version 1.5.2) ▾ Options

Input

Query sRNA in FASTA: No fasta dataset available.

NCBI-nt database: ● NCBI NT (22 Jan 2018) ○ NCBI NT (02 Oct 2015)

Search Parameters

Choose taxon: global

Parameter Setup: ● automatic ○ manual

Additional Settings

Include upstream region: 0

✔ Execute

Fig. 1. Galaxy integration scheme of GLASSgo 1.5.2. The sources on GitHub are used to build a Docker container including all dependencies, which is available on Docker Hub. The Galaxy wrapper defines the user interface and manages the interaction with the Docker container. Special lookup tables required for clade-specific searches can be directly downloaded from Zenodo and are incorporated into Galaxy

2 Results

2.1 Update of GLASSgo

With GLASSgo 1.5.2, we add the possibility to report upstream regions of the found sRNA homologs. This is a prerequisite to infer possible promoter elements of the sRNAs genes, hence facilitating the integration of sRNAs into transcription factor-based gene regulatory networks. To secure compatibility with the latest NCBI databases, GLASSgo 1.5.2 now supports the use of NCBI accession numbers (ACC numbers) as unique identifiers. This adaptation requires the preparation of new lookup tables for the taxonomic classification, which also serve as a prerequisite for clade-specific searches that are often more powerful than unrestricted analyses. To ensure easy updating and retrieval for existing and new installations, these lookup tables have been stored in an open access repository (Zenodo: https://zenodo.org/record/1320180).

2.2 Galaxy integration

The integration of GLASSgo into Galaxy is based on a Docker container. Therein we provide GLASSgo together with all its dependencies (NCBI-BLAST+, RNApdist, …) and distribute it via Docker Hub. Upon new releases the container is built automatically from source. This includes repository tests of the build process itself and functional tests of GLASSgo for different use cases. This Docker container can also be used for command-line-based analyses or for the integration into custom analysis pipelines. For the Galaxy integration, we make use of the Galaxy ToolShed (Blankenberg *et al.*, 2014), which is a one-click solution for tool installation on custom hosted Galaxy instances. GLASSgo follows best practice guidelines and can be seamlessly installed from the Galaxy ToolShed (https://toolshed.g2.bx.psu.edu/view/computationaltranscriptomics/glassgo) using the admin interface. The installation procedure includes functional tests to ensure correct installation and comprehensive documentation of the usage of GLASSgo. Clade-specific searches require some additional configuration that is simplified by the included custom scripts. The automatic interplay of GLASSgo with the local Galaxy environment and external web resources for the Docker container and the lookup tables, which are hosted on Zenodo, is shown in Figure 1. We provide instruction videos that guide through the installation and setup process (https://youtu.be/SiS2ThYDkdU) as well as its usage (https://youtu.be/wFE7LFG9clQ) (Schäfer *et al.*, 2020).

2.3 Using GLASSgo in Galaxy

GLASSgo is part of the RNA workbench (Fallmann *et al.*, 2019), which provides a public Galaxy instance with a set of tools for RNA-related tasks and is available at https://rna.usegalaxy.eu. However, the following description fits also for installations on local Galaxy instances. The user interface follows the design of the GLASSgo web server and is shown in Figure 1. The sRNA sequence of interest has to be uploaded to the user's history in FASTA format. GLASSgo relies on BLAST and, thus, a fundamental parameter is the database to search in. Most Galaxy instances will have a set of databases already available for standard BLAST searches, and GLASSgo can use the same databases for its tasks. If the user wants to use a specialized database, e.g. a clade-specific or a custom database, GLASSgo within Galaxy offers two options: First, the user can choose a clade to restrict the BLAST searches to, which is achieved with the aforementioned lookup tables. Second, users can use custom BLAST databases, for example created from sequences in their own Galaxy history. The usage of GLASSgo is shown in detail in the instruction video mentioned above.

3 Discussion

The integration of GLASSgo into the Galaxy workflow management system offers many advantages over the standalone and web server version such as parameter tracking, version control, batch processing and pipeline development. Galaxy also follows the FAIR manifesto, which stands for 'Findable, Accessible, Interoperable and Reusable' (Wilkinson *et al.*, 2016) and thereby ensures good scientific practice. GLASSgo can be easily installed into a running Galaxy instance through the ToolShed system. In addition, GLASSgo can now be integrated into larger workflows, for example to build RNA family models. Here, it delivers the set of homologous sRNA sequences, and Infernal (Nawrocki *et al.*, 2013) will be used to build the covariance model. In this regard, the incorporation of GLASSgo into Galaxy leverages its full potential with respect to workflow integration and increased usability. We decided on a Docker-based integration because this avoids dependency, compatibility and compilation issues, which frequently occur with other ways to distribute software. Finally, the new feature to include upstream sequences in the results enables new analyses, such as the search for conserved motifs in promoter regions. These can then be used for further validation and filtering, but most importantly allow to integrate sRNAs into gene regulatory networks. Together with recently available advanced tools for sRNA target predictions, this improvement represents another corner stone for the integration of transcription factor-based gene regulatory networks with the post-transcriptional targets of bacterial sRNAs.

Funding

This work was supported by the Federal Ministry of Education and Research (BMBF) programs RNAProNet [031L0106B to W.R.H. and 031L0164A to B.V.], inteRNAct [031A310 to B.V.], the German-Israeli Foundation (GIF) [G-1311 to W.R.H.] and the Deutsche Forschungsgemeinschaft [GE 3159/1-1 to J.G.].

Conflict of Interest: none declared.

References

Afgan,E. *et al.* (2018) The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Res.*, **46**, W537–W544.

Blankenberg,D. *et al.* (2014) Dissemination of scientific software with Galaxy ToolShed. *Genome Biol.*, **15**, 403.

Fallmann,J. *et al*. (2019) The RNA workbench 2.0: next generation RNA data analysis. *Nucleic Acids Res*., **47**, W511–W515.

Lott,S.C. *et al*. (2018) GLASSgo – automated and reliable detection of sRNA homologs from a single input sequence. *Front. Genet*., **9**, 124.

NawrockiE.P. et al (2013) Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics*, **29**, 2933–2935.

Raden,M. *et al*. (2018) Freiburg RNA tools: a central online resource for RNA-focused research and teaching. *Nucleic Acids Res*., **46**, W25–W29.

Schäfer,R.A. *et al*. (2020) GLASSGO Setup & Usage. DaRUS, V2. doi: 10.18419/darus-517.

Wilkinson,M.D. *et al*. (2016) The fair guiding principles for scientific data management and stewardship. *Sci. Data*, **3, 160018.**

Wright,P.R. *et al*. (2018) Workflow for a computational analysis of an sRNA candidate in bacteria. *Methods Mol. Biol*., **1737**, 3–30.