Cell Systems

Report
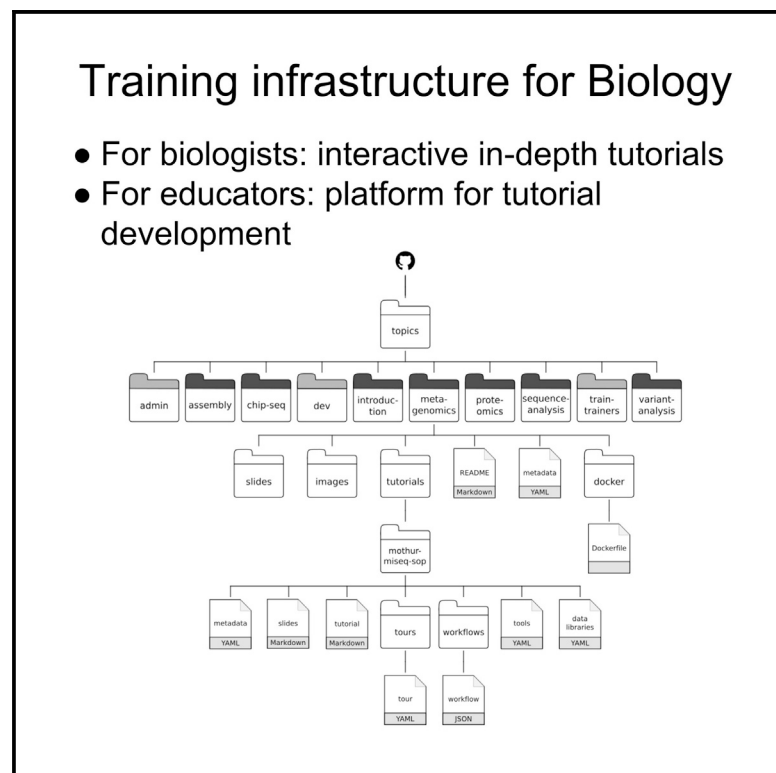
Community-Driven Data Analysis Training for Biology

Graphical Abstract

Highlights

- Web-based, community-maintained infrastructure for data analysis training

- Incorporates latest tools and types of biomedical data

- Supports trainers as much as it supports trainees

Authors

Bérénice Batut, Saskia Hiltemann, Andrea Bagnacani, ..., Rolf Backofen, Anton Nekrutenko, Björn Grüning

Correspondence

backofen@informatik.uni-freiburg.de (R.B.),
anton@nekrut.org (A.N.),
gruening@informatik.
uni-freiburg.de (B.G.)

In Brief

We developed an infrastructure that facilitates data analysis training in life sciences. It is an interactive learning platform tuned for current types of data and research problems. Importantly, it provides a means for community-wide content creation and maintenance and, finally, enables trainers and trainees to use the tutorials in a variety of situations, such as those where reliable Internet access is unavailable.

CellPress

Community-Driven Data Analysis Training for Biology

Bérénice Batut,[1,19] Saskia Hiltemann,[2,19] Andrea Bagnacani,[3] Dannon Baker,[4] Vivek Bhardwaj,[5] Clemens Blank,[1] Anthony Bretaudeau,[6] Loraine Brillet-Guéguen,[7] Martin Čech,[8] John Chilton,[8] Dave Clements,[4] Olivia Doppelt-Azeroual,[9] Anika Erxleben,[1] Mallory Ann Freeberg,[10] Simon Gladman,[11] Youri Hoogstrate,[2] Hans-Rudolf Hotz,[12] Torsten Houwaart,[1] Pratik Jagtap,[13] Delphine Larivière,[8] Gildas Le Corguillé,[14] Thomas Manke,[15] Fabien Mareuil,[9] Fidel Ramírez,[15] Devon Ryan,[15] Florian Christoph Sigloch,[1] Nicola Soranzo,[16] Joachim Wolff,[1] Pavankumar Videm,[1] Markus Wolfien,[3] Aisanjiang Wubuli,[17] Dilmurat Yusuf,[1] Galaxy Training Network,[18] James Taylor,[4] Rolf Backofen,[1,*] Anton Nekrutenko,[8,20,*] and Björn Grüning[1,*]

[1]Bioinformatics Group, Department of Computer Science, Albert-Ludwigs-University Freiburg, Georges-Köhler-Allee 106, Freiburg 79110, Germany
[2]Erasmus Medical Centre, Wytemaweg 80, Rotterdam 3015 CN, the Netherlands
[3]Department of Systems Biology and Bioinformatics, University of Rostock, Ulmenstraße 69, Rostock 18051, Germany
[4]Johns Hopkins University, 3400 N Charles Street, Mudd Hall 144, Baltimore 21218, MD, USA
[5]Department of Biology, Albert-Ludwigs-University, Schänzlestraße 1, Freiburg 79104, Germany
[6]INRA, UMR IGEPP, BIPAA/GenOuest, INRIA/Irisa - Campus de Beaulieu, 35042 RENNES Cedex, France
[7]CNRS, UMPC, FR2424, ABiMS, Station Biologique, Roscoff, France
[8]The Pennsylvania State University, 505 Wartik Lab, University Park, PA 16802, USA
[9]Bioinformatics and Biostatistics HUB, Centre de Bioinformatique, Biostatistique et Biologie Intégrative (C3BI, USR 3756 Institut Pasteur et CNRS), Institut Pasteur, 25-28 Rue du Docteur Roux, 75015 Paris, France
[10]European Bioinformatics Institute, Hinxton, Cambridge, UK
[11]Melbourne Bioinformatics, The University of Melbourne, Melbourne, VIC 3010, Australia
[12]Friedrich Miescher Institute for Biomedical Research, Maulbeerstrasse 66, Basel 4058, Switzerland
[13]Biochemistry, Molecular Biology and Biophysics, University of Minnesota Medical School, 420 Delaware Street SE, Minneapolis, MN 55455, USA
[14]PMC, CNRS, FR2424, ABiMS, Station Biologique, Place Georges Teissier, Roscoff 29680, France
[15]Max Planck Institute of Immunobiology and Epigenetics, Stübeweg 51, Freiburg 79108, Germany
[16]Earlham Institute, Norwich Research Park, Norwich NR4 7UZ, UK
[17]Leibniz Institute for Farm Animal Biology (FBN), Wilhelm-Stahl-Allee 2, Dummerstorf 18196, Germany
[18]https://galaxyproject.org/teach/gtn/
[19]These authors contributed equally
[20]Lead Contact
*Correspondence: backofen@informatik.uni-freiburg.de (R.B.), anton@nekrut.org (A.N.), gruening@informatik.uni-freiburg.de (B.G.)
https://doi.org/10.1016/j.cels.2018.05.012

SUMMARY

The primary problem with the explosion of biomedical datasets is not the data, not computational resources, and not the required storage space, but the general lack of trained and skilled researchers to manipulate and analyze these data. Eliminating this problem requires development of comprehensive educational resources. Here we present a community-driven framework that enables modern, interactive teaching of data analytics in life sciences and facilitates the development of training materials. The key feature of our system is that it is not a static but a continuously improved collection of tutorials. By coupling tutorials with a web-based analysis framework, biomedical researchers can learn by performing computation themselves through a web browser without the need to install software or search for example datasets. Our ultimate goal is to expand the breadth of training materials to include funda-mental statistical and data science topics and to pre-cipitate a complete re-engineering of undergraduate and graduate curricula in life sciences. This project is accessible at https://training.galaxyproject.org.

INTRODUCTION

Rapid development of DNA-sequencing technologies has made it possible for biomedical disciplines to rival the physical sciences in data production capability. The combined output of today's genomics studies has already surpassed the data acquisition rate of entire scientific domains such as astronomy or Internet platforms such as YouTube or Twitter (Stephens et al., 2015). Yet biology is different from astronomy (and other quantitative disciplines) in one fundamental aspect: the lack of computational and data analysis training in standard biomedical curricula. Many biomedical scientists do not possess the skills to use or even access existing analysis resources. Such paucity of training also negatively affects the ability of biological investigators to collaborate with their statistics and

mathematics counterparts because of the inability to speak each other's language. In addition, an estimated one-third of biomedical researchers do not have access to proper data analysis support (Larcombe et al., 2017). The only way to address these deficiencies is with training. The need for such training cannot be overstated: while the majority (>95%) of researchers work or plan to work with large datasets, most (>65%) possess only minimal bioinformatics skills and are not comfortable with statistical analyses (Larcombe et al., 2017) (Williams and Teal, 2017) (Barone et al., 2017). This overwhelming need drives the demand, which, at present, greatly exceeds supply (Attwood et al., 2017). In a recent survey (Community Survey Report, 2013), over 60% of biologists expressed a need for more training, while only 5% called for more computing power. Thus one can assume that the true bottleneck of the current data deluge is not storage or processing power but the knowledge and skills to utilize already existing resources and tools. It is necessary to point out that there are great existing sources of training, such as online teaching materials provided by Johns Hopkins, University of Utah, Rosalind, and others. These are valuable entry points to the field of biomedical data analysis. The type of learning we are describing in this report is complementary to these resources and provides an interactive environment allowing researchers to learn and "play" using pre-configured, data, tools, and computational resources. Importantly, our approach is community driven and thus does not rely on a particular principal investigator, research group, or institution making it potentially more robust and sustainable.

Since 2006, our team has been pondering the question of how to enable computationally naive users to perform complex data analysis tasks. We attempted to solve this problem by creating a platform, Galaxy (http://galaxyproject.org; Afgan et al., 2016), that provides access to hundreds of tools used in a wide variety of analysis scenarios. It features a web-based user interface while automatically and transparently managing underlying computation details. It can be deployed on a personal computer, heterogeneous computer clusters, as well as computation systems provided by Amazon, Microsoft, Google, and other clouds, such as those running OpenStack. Over the years, a community has formed around this project, providing it with an ever-growing, up-to-date set of analysis tools and expanding it beyond life sciences.

These features of Galaxy attracted many biomedical researchers, making it well suited for use as a teaching platform. Here we describe a community-driven effort to build, maintain, and promote a training infrastructure designed to provide computational data analysis training to biomedical researchers worldwide.

RESULTS AND DISCUSSION

Our goal is to develop an infrastructure that facilitates data analysis training in life sciences. At a minimum, it needs to provide an interactive learning platform tuned for current datasets and research problems. It should also provide means for community-wide content creation and maintenance, and, finally, enable trainers and trainees to use the tutorials in a variety of situations, such as those where a reliable Internet access is not an option.

Interactive Learning Tailored to Research Problems

We produced a collection of hands-on tutorials that are designed to be interactive and are built around Galaxy. The hands-on nature of our training material requires that a trainee has two web-browser windows open side by side: one pointed at the current tutorial and the other at a Galaxy instance. We build most tutorials around a "research story": a scenario inspired by a previously published manuscript or an interesting dataset (with the caveat that some more technical materials do not lend themselves to this goal). To make training comprehensive, we aim to cover major branches of biomedical big-data applications, such as those listed in Table 1. Please note that, while we are using Galaxy as an analysis platform, it is not the only way to analyze biomedical data. Thus we design tutorials to teach underlying concepts that will be useful outside Galaxy.

As an example, suppose that a researcher is interested in learning about metagenomic data analyses. The category "Metagenomics" at https://training.galaxyproject.org presently contains a set of introductory slides, two hands-on tutorials, and HTML-based slides designed as a brief (10–20 min) introduction to the subject. In addition, every hands-on tutorial contains background information and explains how it influences data analysis (e.g., Figure 1). This background story is included to account for situations when tutorials are used for self-teaching in the absence of an instructor who would provide a formal introduction. After the introduction, the hands-on part of the tutorial begins and is laid out in a step-by-step fashion with explanations (boxes in Figure 1) of what is being done inside Galaxy, which parameters are critical, and how modifying parameters affects downstream results. The first step in this progression is usually a description of the datasets and how to obtain them. We invested a large effort in creating appropriate datasets by downsampling original published data, which is necessary since real-world datasets are usually too big for tutorials. Our goal was to make datasets as small as possible while still producing an interpretable result. We use Zenodo (http://www.zenodo.org), an open data archiving and distribution platform, to store the tutorial datasets and to provide them with stable digital object identifiers (DOIs) that can be used to credit their authors and for citation purposes.

Tutorials start with a list of prerequisites (typically other tutorials within the site) to account for the variation in trainees' backgrounds, a rough time estimate, questions addressed during the tutorial, learning objectives, and key points. These components help trainees and instructors to keep track of the training goals. For example, the learning objectives are single sentences describing what a trainee will be able to do as a result of the training (Via et al., 2013). Throughout the tutorials, question boxes (Figure 1) are added as an effective way to motivate the trainees (Dollar et al., 2007; Scheines et al., 2005) and guide self-training. The training material is distributed under a Creative Commons BY 4.0 (https://creativecommons.org/licenses/by/4.0/) license: its contents can be shared and adapted freely as long as appropriate credit is given. Efforts have been made also in the direction of ensuring website accessibility to disabled persons by regular evaluation with WAVE (http://wave.webaim.org), a web accessibility evaluation tool, and by automatic checking for alternative text for the images.

Table 1. Topics Available in the Galaxy Training Material Website (https://training.galaxyproject.org) with Their Target Users and Available Tutorials

Topic	Target	Tutorials
Galaxy Server administration	Admin	Galaxy database schema; Docker and Galaxy; advanced customization of a Galaxy instance
Assembly	Biol	Introduction to genome assembly, De Bruijn graph assembly, Unicycler assembly
ChIP-seq data analysis	Biol	Identification of the binding sites of the T cell acute lymphocytic leukemia protein 1, identification of the binding sites of the estrogen receptor
Development in Galaxy	Dev	Contributing with GitHub, tool development and integration into Galaxy, Tool Shed: sharing Galaxy tools, Galaxy interactive tours, Galaxy interactive environments, visualizations: charts plugins, Galaxy Webhooks, visualizations: generic plugins, BioBlend module, a Python library to use Galaxy API, tool dependencies and Conda, tool dependencies and containers, Galaxy code architecture
Epigenetics	Biol	DNA methylation
Introduction to Galaxy	Biol	Galaxy 101, from peaks to genes, multisample analysis, options for using Galaxy, IGV introduction, getting data into Galaxy
Metagenomics	Biol	16S microbial analysis with mothur, analyses of metagenomics data - the global picture
Proteomics	Biol	Protein FASTA database handling, metaproteomics tutorial, label-free versus labelled - how to choose your quantitation method, detection and quantitation of N termini via N-TAILS, peptide and protein ID, secretome prediction, peptide and protein quantification via stable isotope labeling
Sequence Analysis	Biol	Quality control, mapping, genome annotation, RAD-seq reference-based data analysis, RAD-seq de novo data analysis, RAD-seq to construct genetic maps
Train the trainers	Inst	Creating a new tutorial - writing content in Markdown; creating a new tutorial - defining metadata; creating a new tutorial - setting up the infrastructure; creating a new tutorial - creating Interactive Galaxy Tours; creating a new tutorial - building a Docker flavor for a tutorial; good practices to run a workshop

Table 1. *Continued*

Topic	Target	Tutorials
Transcriptomics	Biol	De novo transcriptome reconstruction with RNA-seq, reference-based RNA-seq data analysis, differential abundance testing of small RNAs

Admin, Galaxy administrators; Biol, biomedical researchers; Dev, tool and software developers; Inst, instructors and tutorial developers; RAD-seq, restriction site-associated DNA sequencing; RNA-seq, RNA sequencing. The scripts for the extraction of such information are available in GitHub (https://github.com/bebatut/galaxy-training-material-stats). This table displays content current as of 28th Sep, 2017.
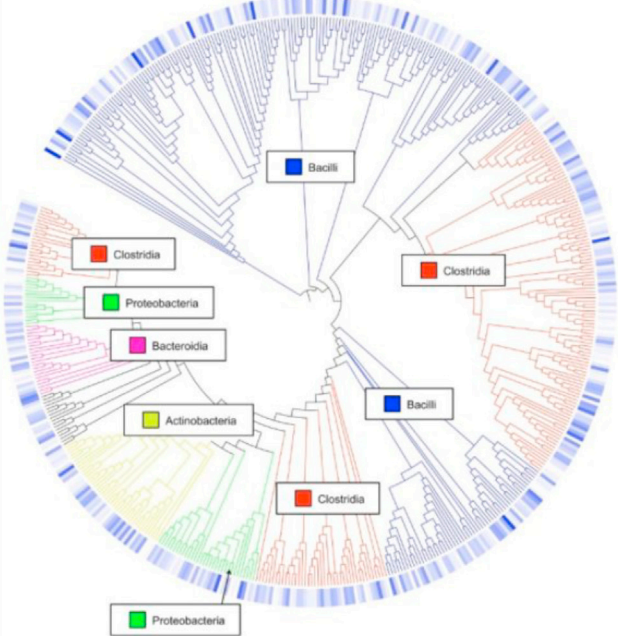
Keeping trainees engaged is critical, particularly for self-training. To this end, we aim to provide interactive tours for each tutorial: using instruction bubbles, each tutorial step can be performed by the user directly inside Galaxy, guiding learners to the needed tools while also allowing exploration of the framework's functionalities. Tours can be created directly in the browser using our tour creator plugin (https://zenodo.org/record/830481).

Infrastructure to Facilitate Community-Led Content Development

To build a comprehensive collection of training materials covering the spectrum of topics in the life sciences, we must leverage community expertise, as no single group can possibly know it all. To achieve this goal, we built an infrastructure that makes tutorial creation a convenient, hassle-free process and enables transparent peer-review and curation to guarantee high-quality and current content. In implementing these requirements, we took inspiration from the Software and Data Carpentry (SDC) projects (Wilson, 2014). In SDC, materials are openly reviewed and iteratively developed on GitHub (https://github.com/) to capture the breadth of community expertise. SDC delivers training via online tutorials with hands-on sections, which offer better training support than videos because trainees who are actively participating learn more (Dollar et al., 2007). This format is also adapted to face-to-face courses and self-training, as the content is openly accessible online. The content of these web pages is easy to edit, thus reducing the contribution barrier. The tutorials are developed in Markdown, a plain text markup language, which is automatically transformed into web-browser-accessible pages. Using these strategies, we created a GitHub repository (https://github.com/galaxyproject/training-material) to collect, manage, and distribute training materials. The architecture of this infrastructure is shown in Figure 2 (center), with the process for developing a tutorial illustrated at the bottom of the figure. To create a new tutorial, the main repository is "forked" (duplicated into a user-controlled space) within GitHub by an individual developing the tutorial. The developer then proceeds to write the content using Markdown, as explained in our guide at https://training.galaxyproject.org/topics/contributing (itself consisting of several tutorials). The guide contains detailed information on technical and stylistic aspects of tutorial development. After settling on a final version of the tutorial (circles 1–10, bottom of Figure 2), a "pull request" is created against the original repository. When a new pull request is issued, this is an indication that a new tutorial is ready to be

A Background: Operational Taxonomic Units (OTUs)

In 16S metagenomics approaches, OTUs are clusters of similar sequence variants of the 16S rDNA marker gene sequence. Each of these clusters is intended to represent a taxonomic unit of a bacteria species or genus depending on the sequence similarity threshold. Typically, OTU cluster are defined by a 97% identity threshold of the 16S gene sequence variants at species level. 98% or 99% identity is suggested for strain separation.

(Image credit: Danzeisen et al. 2013, 10.7717/peerj.237)

B Hands-on: Cluster mock sequences into OTUs

First we calculate the pairwise distances between our sequences
- **Dist.seqs** 🔧 with the following parameters
    - "fasta" to the fasta from Get.groups
    - "cutoff" to 0.20

Next we group sequences into OTUs
- **Cluster** 🔧 with the following parameters
    - "column" to the dist output from Dist.seqs
    - "count" to the count table from Get.groups

Now we make a *shared* file that summarizes all our data into one handy table
- **Make.shared** 🔧 with the following parameters
    - "list" to the OTU list from Cluster
    - "count" to the count table from Get.groups
    - "label" to 0.03 (this indicates we are interested in the clustering at a 97% identity threshold)

And now we generate intra-sample rarefaction curves
- **Rarefaction.single** 🔧 with the following parameters
    - "shared" to the shared file from Make.shared

Question

How many OTUs were identified in our mock community?
▶ Click to view answer

Figure 1. Key Elements of an Interactive Tutorial
(A) A fragment of introductory material within a tutorial.
(B) A "hands-on" element in the upper box contains instructions for running a tool inside Galaxy. The question box at the bottom contains an answer field that can be toggled.
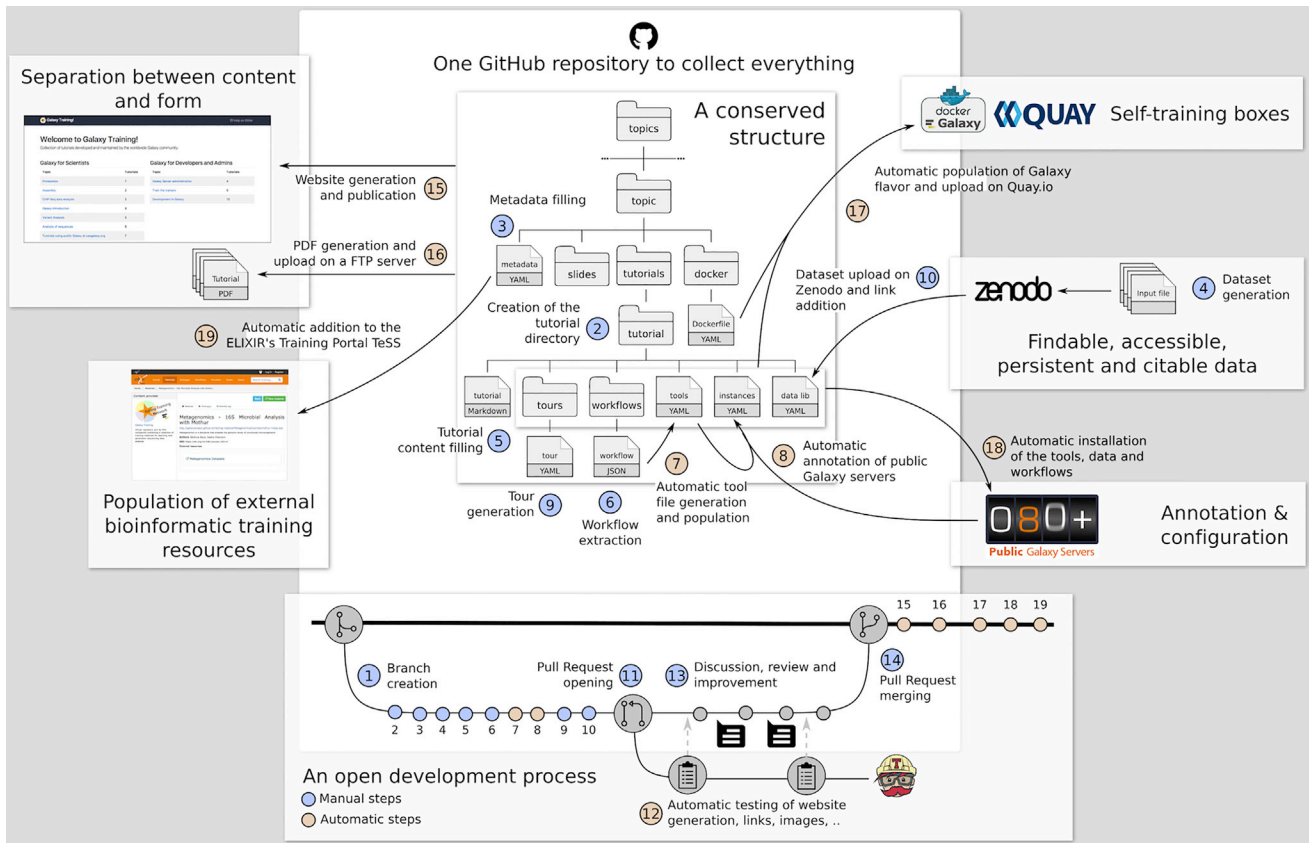
Figure 2. Structure and Development of Content in GitHub (http://github.com/galaxyproject/training-material)
The material is organized in different topics, each topic in a dedicated directory. Inside each topic's directory, the structure is the same: a metadata file, a directory with the topic introduction slide decks, a directory with the tutorials, and a directory with the Dockerfile describing the details to build a container for the topic that would contain a dedicated Galaxy instance with all tools relevant for the tutorials. Inside the topic directory, each tutorial related to the topic has its own subdirectory with several files: a tutorial file written in Markdown with hands-on, an optional slides file to support the tutorial, a directory with Galaxy interactive tours to reproduce the tutorial, a directory with workflows extracted from the tutorial, a file with the links to the input data needed for the tutorial, and a file with the description of needed tools to run the tutorial. The process of development of new content is shown at the bottom.

reviewed by the editorial team. The team then makes suggestions on the new contents, these suggestions are discussed, and the content is edited accordingly. A decision is then made whether to accept the pull request. At the same time the pull request is first created, the newly added content is automatically tested for HTML generation and all links and images are verified. When the pull request is accepted, the new tutorial becomes a part of the official training material portfolio, and the entire site is regenerated. This open strategy for content creation started paying off early as we already have over 60 individuals contributing and editing content within the GitHub repository.

This infrastructure has been developed in accordance with the FAIR (findable, accessible, interoperable, reusable) principles (Wilkinson et al., 2016). Each tutorial, slide deck, and topic is complemented by numerous metadata described in a standard, accessible, interoperable format (YAML; http://yaml.org/). The metadata are used to automatically populate the TeSS training portal at the European Life Sciences Infrastructure for Biological Information (ELIXIR; https://tess.elixir-europe.org), ensuring global reach (Beard et al., 2016). Each topic, tutorial, and slide deck has as metadata a reference to a topic in the EDAM

ontology (Ison et al., 2013), a comprehensive catalog of well-established, familiar concepts that are prevalent within bioinformatics and computational biology. These references can be used to represent relationships among the materials and make them more findable and searchable.

Using the framework described above, we relaunched the Galaxy Training Network (GTN; https://galaxyproject.org/teach/gtn). This growing network currently consists of 33 scientific groups (https://galaxyproject.org/teach/trainers) invested in Galaxy-based training. The GTN regularly organizes training events worldwide (Figure S1) and offers best practices for developing Galaxy-based training material, advice on computer platform choice to use for training, and a catalog of existing training resources for Galaxy (Table 1).

As of writing (March 2018) 64 individuals contributed to development of infrastructure and tutorials (http://bit.ly/gxy_tr_people). Of these, only 18 individuals are associated with the two largest Galaxy Project installations in the United States (http://usegalaxy.org) and Germany (http://usegalaxy.eu). This ratio ($46/18 \approx 2.5$) is an indicator of community engagement and our goal is to increase it.

Ensuring Accessibility of Tutorials

Most training materials hosted within the GTN resource are intended to be used side by side with the Galaxy framework. However, the main public Galaxy instances (e.g., https://usegalaxy.org or https://usegalaxy.eu) are occasionally subject to unpredictable load, may be inaccessible due to network problems in remote parts of the world, or may not have all the tools necessary for completing the tutorials. To account for these situations, we have developed a Docker-based framework for creating portable, on-demand Galaxy instances specifically targeted for a given tutorial. Docker (https://www.docker.com) is a container platform that provides lightweight virtualization by executing "images" (files that include everything needed to run a piece of software) isolated from the host computer environment. An individual creating a new tutorial lists all tools that are required to complete it in a dedicated configuration file (tools file, Figure 2). For example, a metagenomics tutorial uses the mothur (Schloss et al., 2009) set of tools as well as visualization applications such as Krona (Ondov et al., 2015). The corresponding Galaxy tools are listed in a configuration file that is a part of the metagenomics tutorial. This file is used to install these Galaxy tools and their dependencies into a base Galaxy Docker image (containing essential Galaxy functionality and a core set of tools) to create a dedicated "on-demand" Galaxy instance that can then be used on any trainer's or trainee's computer. The Docker image also contains input data, tours, and workflows.

A Vision for the Future

Life sciences are on a trajectory toward becoming an entirely data-driven scientific domain. A growing understanding that biomedical curricula must be modernized to reflect these changes is gaining attention (Hitchcock et al., 2017). Our project represents one of the first fully open, "grass-roots" attempts at unifying and standardizing heterogeneous training resources around the Galaxy platform. While it may not be appropriate to all, our multi-year experience with teaching workshops at various skill levels can be summarized as the following set of recommendations, which we use as guiding principles. These recommendations may also be useful for the development of alternative frameworks as well as for curriculum planning:

1. Require quantitative training. No one expects biomedical researchers to rival their colleagues in departments of mathematics or statistics. However, background level statistical reasoning must be included in all training materials and general statistical courses must become a part of undergraduate and graduate education. This would have an enormous positive impact on the quality of biomedical research because researchers with basic understanding of quantitative concepts will not, for example, perform an RNA sequencing experiment without a sufficient number of replicates. While our current set of tutorials lacks in-depth statistical analyses of the data, we are planning to change this. Our integration with Jupyter is the first step in this direction (Grüning et al., 2017).

2. Demystify computational methodologies. Fundamental principles, limitations, and assumptions of molecular experimental techniques are typically well understood by biomedical researchers even when proprietary reagent kits are used. This is not the case with software tools, which are often treated as black boxes. We argue that fundamental principles of bioinformatic techniques (e.g., read mapping, read assembly) must be understood by experimentalists as this will also lead to an increase in overall quality of research output.

3. Advocate the fundamental virtues of open and transparent research. Open and transparent data analysis (e.g., through the use of open-source software) promotes replication and validation of results by independent investigators. It also speeds up research progress by facilitating reuse and repurposing of published analyses to different datasets or even to other disciplines. We advocate openness as a basic principle for computational analysis of biomedical data.

The infrastructure presented here has been developed to support training using Galaxy, a powerful tool for teaching bioinformatics concepts and analysis, but such a model is not only limited to Galaxy. It could be applied to bioinformatics training more generally (and to other disciplines as well) to support learners and instructors in this ever-changing landscape that is the life sciences.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- CONTACT FOR REAGENT AND RESOURCE SHARING
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
- METHOD DETAILS
- QUANTIFICATION AND STATISTICAL ANALYSIS
- DATA AND SOFTWARE AVAILABILITY

AUTHOR CONTRIBUTIONS

B.B., S.H., and B.G. developed the conceptual foundation for the training infrastructure, developed the proof of principle, and outlined the software process. B.B., S.H., R.B., and A.N. wrote the manuscript. A. Bagnacani, D.B., V.B., C.B., A. Bretaudeau, L.B.-G., M.Č., J.C., D.C., G.T.N., O.D.-A., A.E., M.A.F., S.G., Y.H., H.-R.H., T.H., P.J., D.L., G.L.C., T.M., F.M., F.R., D.R., F.C.S., N.S., J.T., J.W., P.V., M.W., A.W., and D.Y. contributed software components and tutorials and also edited and commented on the manuscript.

REFERENCES

Afgan, E., Baker, D., van den Beek, M., Blankenberg, D., Bouvier, D., Čech, M., Chilton, J., Clements, D., Coraor, N., Eberhard, C., et al. (2016). The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. Nucleic Acids Res. 44, W3–W10.

Attwood, T.K., Blackford, S., Brazas, M.D., Davies, A., and Schneider, M.V. (2017). A global perspective on evolving bioinformatics and data science training needs. Brief. Bioinform. https://doi.org/10.1093/bib/bbx100.

Barone, L., Williams, J., and Micklos, D. (2017). Unmet needs for analyzing biological big data: a survey of 704 NSF principal investigators. PLoS Comput. Biol. 13, e1005755.

Beard, N., Attwood, T., and Nenadic, A. (2016). TeSS – training portal. F1000Res. 5, https://doi.org/10.7490/f1000research.1112652.1.

Community Survey Report – 2013. EMBL Australia Bioinformatics Resource. https://www.embl-abr.org.au/news/braembl-community-survey-report-2013/.

Dollar, A., Steif, P.S., and Strader, R. (2007). Enhancing traditional classroom instruction with web-based Statics course. In: 2007 37th Annual Frontiers in Education Conference - Global Engineering: Knowledge without Borders, Opportunities without Passports.

Grüning, B.A., Rasche, E., Rebolledo-Jaramillo, B., Eberhard, C., Houwaart, T., Chilton, J., Coraor, N., Backofen, R., Taylor, J., and Nekrutenko, A. (2017). Jupyter and Galaxy: easing entry barriers into complex data analyses for biomedical researchers. PLoS Comput. Biol. 13, e1005425.

Hitchcock, P., Mathur, A., Bennett, J., Cameron, P., Chow, C., Clifford, P., Duvoisin, R., Feig, A., Finneran, K., Klotz, D.M., et al. (2017). The future of graduate and postdoctoral training in the biosciences. Elife 6, https://doi.org/10.7554/eLife.32715.

Ison, J., Kalas, M., Jonassen, I., Bolser, D., Uludag, M., McWilliam, H., Malone, J., Lopez, R., Pettifer, S., and Rice, P. (2013). EDAM: an ontology of bioinformatics operations, types of data and identifiers, topics and formats. Bioinformatics 29, 1325–1332.

Larcombe, L., Hendricusdottir, R., Attwood, T.K., Bacall, F., Beard, N., Bellis, L.J., Dunn, W.B., Hancock, J.M., Nenadic, A., Orengo, C., et al. (2017). ELIXIR-UK role in bioinformatics training at the national level and across ELIXIR. F1000Res. 6, https://doi.org/10.12688/f1000research.11837.1.

Ondov, B.D., Bergman, N.H., and Phillippy, A.M. (2015). Krona: interactive metagenomic visualization in a web browser. In Encyclopedia of Metagenomics, K.E. Nelson, ed. (Springer), pp. 339–346.

Scheines, R., Leinhardt, G., Smith, J., and Cho, K. (2005). Replacing lecture with web-based course materials. J. Educ. Comput. Res. 32, 1–25.

Schloss, P.D., Westcott, S.L., Ryabin, T., Hall, J.R., Hartmann, M., Hollister, E.B., Lesniewski, R.A., Oakley, B.B., Parks, D.H., Robinson, C.J., et al. (2009). Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl. Environ. Microbiol. 75, 7537–7541.

Stephens, Z.D., Lee, S.Y., Faghri, F., Campbell, R.H., Zhai, C., Efron, M.J., Iyer, R., Schatz, M.C., Sinha, S., and Robinson, G.E. (2015). Big data: astronomical or genomical? PLoS Biol. 13, e1002195.

Via, A., Blicher, T., Bongcam-Rudloff, E., Brazas, M.D., Brooksbank, C., Budd, A., De Las Rivas, J., Dreyer, J., Fernandes, P.L., van Gelder, C., et al. (2013). Best practices in bioinformatics training for life scientists. Brief. Bioinform. 14, 528–537.

Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L.B., Bourne, P.E., et al. (2016). The FAIR guiding principles for scientific data management and stewardship. Sci. Data 3, 160018.

Williams, J.J., and Teal, T.K. (2017). A vision for collaborative training infrastructure for bioinformatics. Ann. N. Y. Acad. Sci. 1387, 54–60.

Wilson, G. (2014). Software carpentry: lessons learned. F1000Res. 3, 62.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Software and Algorithms		
GitHub software development and distribution platform		github.com
Galaxy Project		galaxyproject.org
Galaxy main public site US		usegalaxy.org
Galaxy main public site EU		usegalaxy.eu
Galaxy Training Materials	This paper	http://github.com/galaxyproject/training-material

CONTACT FOR REAGENT AND RESOURCE SHARING

The Lead Contact is Anton Nekrutenko (anton@nekrut.org). There are no restrictions on the use of reported resources. This resource is licensed under the Creative Commons Attribution 4.0 International License.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

No experiments have been performed in the framework of this study.

METHOD DETAILS

Training materials are developed using Markdown markup language and served from the GitHub platform. Extensive description of the development process and content structure can be found at https://github.com/galaxyproject/training-material/blob/master/CONTRIBUTING.md.

QUANTIFICATION AND STATISTICAL ANALYSIS

No quantification or statistical analyses have been performed in this study.

DATA AND SOFTWARE AVAILABILITY

All tutorials are available from http://galaxyproject.github.io/training-material/.