Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

HHS Public Access

Author manuscript

Nat Methods. Author manuscript; available in PMC 2024 May 05.

Published in final edited form as:

Nat Methods. 2018 July ; 15(7): 475–476. doi:10.1038/s41592-018-0046-7.

Bioconda: sustainable and comprehensive software distribution for the life sciences

Björn Grüning[1,12], **Ryan Dale**[2,12], **Andreas Sjödin**[3,4], **Brad A. Chapman**[5], **Jillian Rowe**[6], **Christopher H. Tomkins-Tinch**[7,8], **Renan Valieris**[9], **Johannes Köster**[10,11,*], **The Bioconda Team**[13]

[1]Bioinformatics Group, Department of Computer Science, University of Freiburg, Freiburg, Germany.

[2]Laboratory of Cellular and Developmental Biology, National Institute of Diabetes and Digestive and Kidney Diseases, US National Institutes of Health, Bethesda, MD, USA.

[3]Division of CBRN Security and Defence, FOI–Swedish Defence Research Agency, Umeå, Sweden.

[4]Department of Chemistry, Computational Life Science Cluster (CLiC), Umeå University, Umeå, Sweden.

[5]Harvard T.H. Chan School of Public Health, Boston, MA, USA.

[6]Center for Genomics and Systems Biology, Genomics Core,, NYU Abu Dhabi,, Abu Dhabi,, United Arab Emirates.

[7]Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA, USA.

[8]Broad Institute of MIT and Harvard, Cambridge, MA, USA.

[9]Laboratory of Bioinformatics and Computational Biology, A. C. Camargo Cancer Center, São Paulo, Brazil.

[10]Algorithms for Reproducible Bioinformatics, Genome Informatics, Institute of Human Genetics, University Hospital Essen, University of Duisburg– Essen, Essen, Germany.

[11]Medical Oncology, Dana Farber Cancer Institute, Harvard Medical School, Boston, MA, USA.

[12]These authors contributed equally: Björn Grüning and Ryan Dale.

[13]A full list of authors and affiliations is available as Supplementary Table 1.

[*] johannes.koester@uni-due.de .

Author contributions

J.K. and R.D. wrote the manuscript and conducted the data analysis. K. Beauchamp, C. Brueffer, B.A.C., F. Eggenhofer, B.G., E. Pruesse, M. Raden, J.R., D. Ryan, I. Shlyakter, A.S., C.H.T.-T., and R.V. (in alphabetical order) contributed to writing of the manuscript. D.A. Søndergaard supervised student programmers on writing Conda package recipes and maintaining the connection with ELIXIR. All other members of the Bioconda Team contributed or maintained recipes (author order was determined by the number of commits in October 2017).

Reporting Summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this article.

Competing interests

The authors declare no competing interests.

Additional Information

Supplementary information is available for this paper at https://doi.org/10.1038/s41592-018-0046-7.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

To the Editor:

Bioinformatics software comes in a variety of programming languages and requires diverse installation methods. This heterogeneity makes management of a software stack complicated, error-prone, and inordinately time-consuming. Whereas software deployment has traditionally been handled by administrators, ensuring the reproducibility of data analyses[1–3] requires that the researcher be able to maintain full control of the software environment, rapidly modify it without administrative privileges, and reproduce the same software stack on different machines.

The Conda package manager (https://conda.io) has become an increasingly popular means to overcome these challenges for all major operating systems. Conda normalizes software installations across language ecosystems by describing each software with a human readable 'recipe' that defines meta-information and dependencies, as well as a simple 'build script' that performs the steps necessary to build and install the software. Conda builds software packages in an isolated environment, transforming them into relocatable binaries. Importantly, it obviates reliance on system-wide administration privileges by allowing users to generate isolated software environments in which they can manage software versions by project, without generating incompatibilities and side effects (Supplementary Results). These environments support reproducibility, as they can be rapidly exchanged via files that describe their installation state. Conda is tightly integrated into popular solutions for reproducible data analysis such as Galaxy[4], bcbio-nextgen (https://github.com/chapmanb/bcbio-nextgen), and Snakemake[5]. To further enhance reproducibility guarantees, Conda can be combined with container or virtual machine-based approaches and archive facilities such as Zenodo (Supplementary Results). Finally, although Conda provides many commonly used packages by default, it also allows users to optionally include additional, community-managed repositories of packages (termed channels).

To unlock the benefits of Conda for the life sciences, we present the Bioconda project (https://bioconda.github.io). The Bioconda project provides over 3,000 Conda software packages for Linux and macOS. Rapid turnaround times (Supplementary Results) and extensive documentation (https://bioconda.github.io/contributing.html) have led to a growing community of over 200 international scientists working in the project (Supplementary Results). The project is led by a core team, which is complemented by interest groups for particular language ecosystems. Unlimited (in time and space) storage for generated packages is donated by Anaconda Inc. All other used infrastructure is free of charge. Bioconda provides packages from various language ecosystems such as Python, R (CRAN and Bioconductor), Perl, Haskell, Java, and C/C++ (Fig. 1a). Many of the packages have complex dependency structures that require various manual steps for installation when not relying on a package manager like Conda (Supplementary Results). With over 6.3 million downloads, Bioconda has become a backbone of bioinformatics infrastructure that is used heavily across all language ecosystems (Fig. 1b). It is complemented by the conda-forge project (https://condaforge.github.io), which hosts software not specifically related to the biological sciences. This separation has proven beneficial, because the focused nature of the Bioconda community allows for fast turnaround times and support when a user needs to

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

contribute packages or fix problems. Nevertheless, the two projects collaborate closely, and the Bioconda team maintains over 500 packages hosted by conda-forge.

Bioconda is not the only effort to distribute bioinformatics software (Fig. 1c). The alternatives can be categorized into system-wide (Debian-Med, Genotoo Science, Biolinux, and Homebrew) and per-user (EasyBuild, GNU Guix, and BioBuilds) installation mechanisms. The system-wide approaches lack the ability to put the scientist in control of the installed software stack, and thus do not meet the requirements for reproducibility outlined above. All per-user-based approaches provide a similar feature set (BioBuilds is also using the Conda package manager). However, among all available approaches, Bioconda, despite being the most recent, is by far the most comprehensive, with thousands of software libraries and tools that are maintained by hundreds of international contributors (Fig. 1c).

For reproducible data science, it is crucial that software libraries and tools be provided via an easy-to-use, unified interface, so that they can be easily deployed and sustainably managed. With its ability to maintain isolated software environments, integration into major workflow management systems, and lack of requirement for any administration privileges for use, the Conda package manager is the ideal tool to ensure sustainable and reproducible software management. Bioconda packages have been well received by the community, with over six million downloads so far. We invite everybody to join the Bioconda community, participate in maintaining or publishing new software, and work toward the goal of a central, comprehensive, and language-agnostic collection of easily installable software for the life sciences.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We thank all contributors, the conda-forge team, and Anaconda Inc. for excellent cooperation. Further, we thank Travis CI (https://travis-ci.com) and Circle CI (https://circleci.com) for providing free Linux and macOS computing capacity. Finally, we thank ELIXIR (https://www.elixir-europe.org) for constant support and donation of staff. This work was supported by the Intramural Program of the National Institute of Diabetes and Digestive and Kidney Diseases, US National Institutes of Health (R.D.), the Netherlands Organisation for Scientific Research (NWO) (VENI grant 016. Veni.173.076 to J.K.), the German Research Foundation (SFB 876 to J.K.), and the NYU Abu Dhabi Research Institute for the NYU Abu Dhabi Center for Genomics and Systems Biology, program number CGSB1 (grant to J.R. and A. Yousif).

Data availability.

Data and code underlying the presented results are enclosed in a Snakemake workflow archive available at https://doi.org/10.5281/zenodo.1068297. The archive can also be used to automatically reproduce all results and figures presented in this paper.

References

1. Mesirov JP Science 327, 415–416 (2010). [PubMed: 20093459]
2. Baker M. Nature 533, 452–454 (2016). [PubMed: 27225100]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

3. Munafò MR et al. Nat. Hum. Behav. 1, 0021 (2017). [PubMed: 33954258]
4. Afgan E. et al. Nucleic Acids Res. 44, W3–W10 (2016). [PubMed: 27137889]
5. Köster J. & Rahmann S. Bioinformatics 28, 2520–2522 (2012). [PubMed: 22908215]
6. Field D. et al. Nat. Biotechnol. 24, 801–803 (2006). [PubMed: 16841067]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

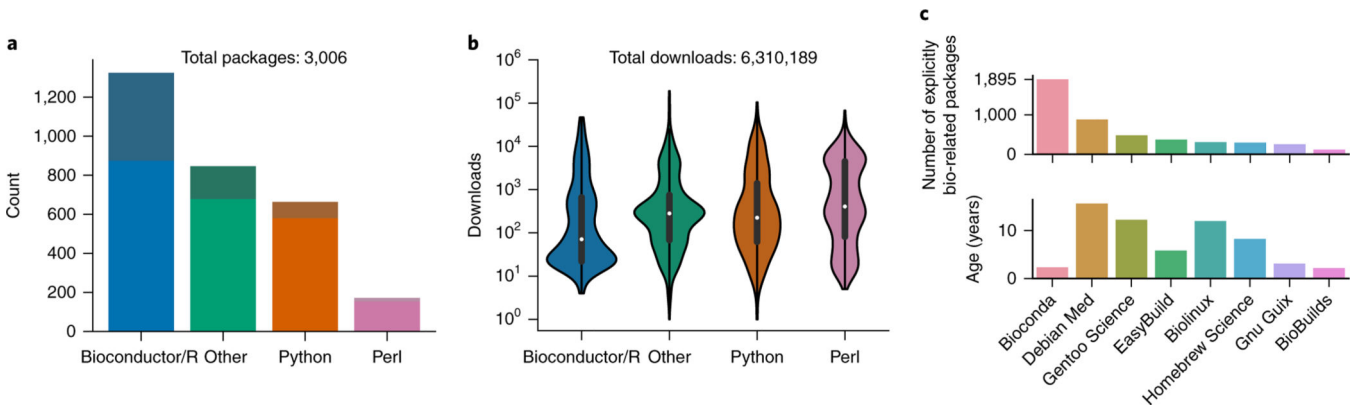

Fig. 1 |. Package numbers and usage.

a, Package count per language ecosystem (saturated colors on the lower portions of the bars represent explicitly life-science-related packages). **b**, Distribution of per-package downloads, separated by language ecosystem. The term “other” encompasses all packages that do not fall into one of the specific categories named. White dots represent the mean; dark bars represent the interval between upper and lower quartiles. **c**, Comparison of the number of explicitly life-science-related packages in Bioconda with that in Debian Med (https://www.debian.org/devel/debian-med), Gentoo Science Overlay (category sci-biology; https://github.com/gentoo/sci), EasyBuild (module bio; https://easybuilders.github.io/easybuild), Biolinux[6], Homebrew Science (tag bioinformatics; https://brew.sh), GNU Guix (category bioinformatics; https://www.gnu.org/s/guix), and BioBuilds (https://biobuilds.org). The lower graph shows the project age since the first release or commit. Statistics obtained 25 October 2017.