Databases and ontologies

BioArchLinux: community-driven fresh reproducible software repository for life sciences

Guoyi Zhang[1,2,*], Pekka Ristola[3], Han Su[4], Bipin Kumar[5], Boyu Zhang[6], Yujin Hu[7], Michael G Elliot[8], Viktor Drobot[9], Jie Zhu[10,11], Jens Staal[12], Martin Larralde[13,14], Shun Wang[15], Yun Yi[4] and Haoran Yu[16]

[1]Evolution & Ecology Research Centre, School of Biological, Earth, Environmental Sciences, UNSW Sydney, Australia, [2]Australian Museum Research Institute, Australian Museum, Sydney, Australia, [3]BioArchLinux member, Helsinki, Finland, [4]BioArchLinux member, Beijing, China, [5]Obront Biotech Pvt. Ltd, Hyderabad, India, [6]Beijing Forestry University, Beijing, China, [7]Shenzhen Research Institute of Big Data, Shenzhen, China, [8]Department of Evolution. University of Groningen, Groningen, Netherlands, [9]Pharm InnTech LLC, Moscow, Russia, [10]Microbiota I-Center (MagIC), Hong Kong SAR, 999077, China, [11]Li Ka Shing Institute of Health Sciences, Faculty of Medicine, The Chinese University of Hong Kong, Hong Kong SAR, 999077, China, [12]Department of Biomedical Molecular Biology, Ghent University, Belgium, [13]Leiden University Center for Infectious Diseases, Leiden University Medical Center (LUMC), Leiden, Netherlands, [14]Structural and Computational Biology Unit, EMBL, Heidelberg, Germany, [15]JASP team member, Shenzhen, China, [16]Xi’an University, Xi’an, China.

*Evolution & Ecology Research Centre, School of Biological, Earth, Environmental Sciences, UNSW Sydney, Australia; Australian Museum Research Institute, Australian Museum, Sydney, Australia. Email: starsareintherose@gmail.com

Associate Editor: Peter Robinson

Received on 19 November 2024; revised on 03 March 2025; accepted on 04 March 2025

Abstract

Motivation: The BioArchLinux project was initiated to address challenges in bioinformatics software reproducibility and freshness. Relying on Arch Linux's user-driven ecosystem, we aim to create a comprehensive and continuously updated repository for life sciences research.

Results: BioArchLinux provides a PKGBUILD-based system for seamless software packaging and maintenance, enabling users to access the latest bioinformatics tools across multiple programming languages. The repository includes Docker images, Windows Subsystem for Linux (WSL) support, and Junest for non-root environments, enhancing accessibility across platforms. Although being developed and maintained by a small core team, BioArchLinux is a fast-growing bioinformatics repository that offers a participatory and community-driven environment.

Availability: The repository, documentation, and tools are freely available at https://bioarchlinux.org and https://github.com/BioArchLinux. Users and developers are encouraged to contribute and expand this open-source initiative.

Contact: starsareintherose@gmail.com

1 Introduction

Bioinformatics software and the development of new algorithms and tools are crucial for analyzing and interpreting biological data (Hettrick *et al.* 2014), enabling researchers to make significant discoveries in genomics, proteomics, and other fields (Horner *et al*. 2010) by facilitating the processing of large datasets and visualization of complex biological processes. One concern with an increased reliance on complex software in data processing is the transparency of the methods used, and the reproducibility of the processes employed for the analysis (Waltemath and Wolkenhauer 2016, Kanterakis *et al.* 2019, Raghupathi *et al*. 2022). Reproducibility ensures that conclusions from scientific experiments and analyses can be trusted, and it is essential for validating results and building upon previous work (Baker 2016). However, when it comes to bioinformatics software, reproducibility is related to a multitude of factors (Baykal *et al*. 2024), in particular the chosen platform.

Linux is a major bioinformatics analysis platform known for its flexibility and modularity (Yoo 2016). However, this also has a problem: distribution community fragmentation. Fragmented communities have different compilation processes, and compilation is an important factor affecting the repeatability of bioinformatics software (Grüning *et al*. 2018). The distribution's "software freshness" (i.e., packages are at the latest official release) affects whether users can use the most up-to-date software and the distribution's software coverage, reflecting the availability of pre-compiled software in repositories. Those two factors will influence whether users choose to compile software by themselves or use pre-compiled software provided by repositories.

Users often choose to compile software themselves when the community-provided software is insufficient or outdated. Facing the significant challenges of bioinformatics software on repeatability and freshness, Arch Linux excels in providing the latest software and an easy-to-use functional shell script packing system, making it an ideal mainstream distribution for freshness (Legay *et al*. 2021, Repology Team 2025) and easy user-driven contribution that ensures a high software coverage (e.g. ArchLinux User Repository is the second largest Linux user-contributed repository based on repology statistics due to Arch Linux package style). Considering these factors, we initiated the BioArchLinux project based on the Arch Linux environment to address these issues to provide a more efficient solution to the problem of software freshness and coverage in bioinformatics.

2 Implementation

BioArchLinux brings packages that are also offered by other bioinformatics repositories, covering a wide range of tools—from traditional morphology analysis software developed over the past 30 years to state-of-the-art genomics analysis tools. Our platform is designed to run natively on x86 bare metal and supports macOS via Docker, Windows via WSL, and Linux HPC environments, where installations can be executed without root privileges using Junest.

We use a PKGBUILD-based system to create and manage packages (Fig. 1A). Shell scripts are familiar and widely used, and their function-based structure allows for a clear definition of different packaging processes such as preparation, compilation, and packaging. Our packaging tool, lilac, can handle lilac.yaml and lilac.py scripts facilitating easy interaction with Python for version updates. This allows a team of around 15 people to maintain over 5,000 packages as well as developing and maintaining development software. Although many of the packages we maintain are R packages, which are relatively easier to manage and which significantly contribute to the growth of our repository, we currently maintain around 500 non-R packages.

Given the complexity of downstream dependencies, we ensure reproducibility by providing container images, such as Docker and Junest, that encapsulate the entire environment, enabling analyses to be consistently replicated. In doing so, our Docker images offer a consistent and isolated platform for running bioinformatics software on macOS, Windows, and other Linux distributions. Fresh Docker images are provided daily to ensure users always have access to the latest software, usually about 200 MB. Junest (Squillace 2024) provides a lightweight, non-root solution for running Arch Linux packages on other Linux distributions, making it accessible to a broader range of non-root users. For Windows users, we support the Windows Subsystem for Linux (WSL), allowing them to run Linux-based bioinformatics tools natively on Windows 10 and above.

Pacman 7.0 introduces new features such as sandboxing and DownloadUser, enhancing security and user experience by isolating package installations and managing downloads more efficiently. Our Git repository also offers excellent tracking capabilities to ensure repeatability.

Although Arch Linux provides a snapshot website to help users downgrade and roll back, we currently lack the resources to offer this functionality ourselves. Users can create the environment using Arch Linux archive snapshots and rebuild packages using our git repository to achieve reproducibility. Once additional computing resources become available, we aim to provide snapshot capabilities to further enhance reproducibility.

For some of our maintained packages, such as beagle-libs, IQ-TREE, MrBayes, and Eugene, we offer various specialized optimization versions of pre-compiled packages. These include optimizations utilizing GPU (OpenCL/CUDA) technologies and parallelization. These optimizations cater to different user needs and deliver enhanced performance. Providing more optimization options to users will always be our future aim.

To better serve the whole Linux software ecology instead of only our maintained repository, our collaboration with upstream and Arch Linux maintainers has been a cornerstone of our efforts, including the regular submission of patches upstream. Given that Arch Linux is an aggressively updated distribution, we are often the first community to identify and address source code and latest version compiler incompatibility issues, for example, GCC. Our contributions span a variety of packages, including RevBayes, TrinityRNAseq, Augustus, eugene, BEAST-mcmc, Open Delta, and more.

In addition to package maintenance, we have developed several tools to aid in the upkeep of Arch Linux-based distributions and projects. We provide a comprehensive package search page to help users find relevant information easily. We have also co-contributed the development of the lilac building tool, which has proven invaluable for other Arch Linux-based distributions. Furthermore, the nvchecker tool, which we use and contribute to, has been widely adopted by many Linux distributions, underscoring our commitment to enhancing the broader Linux ecosystem.

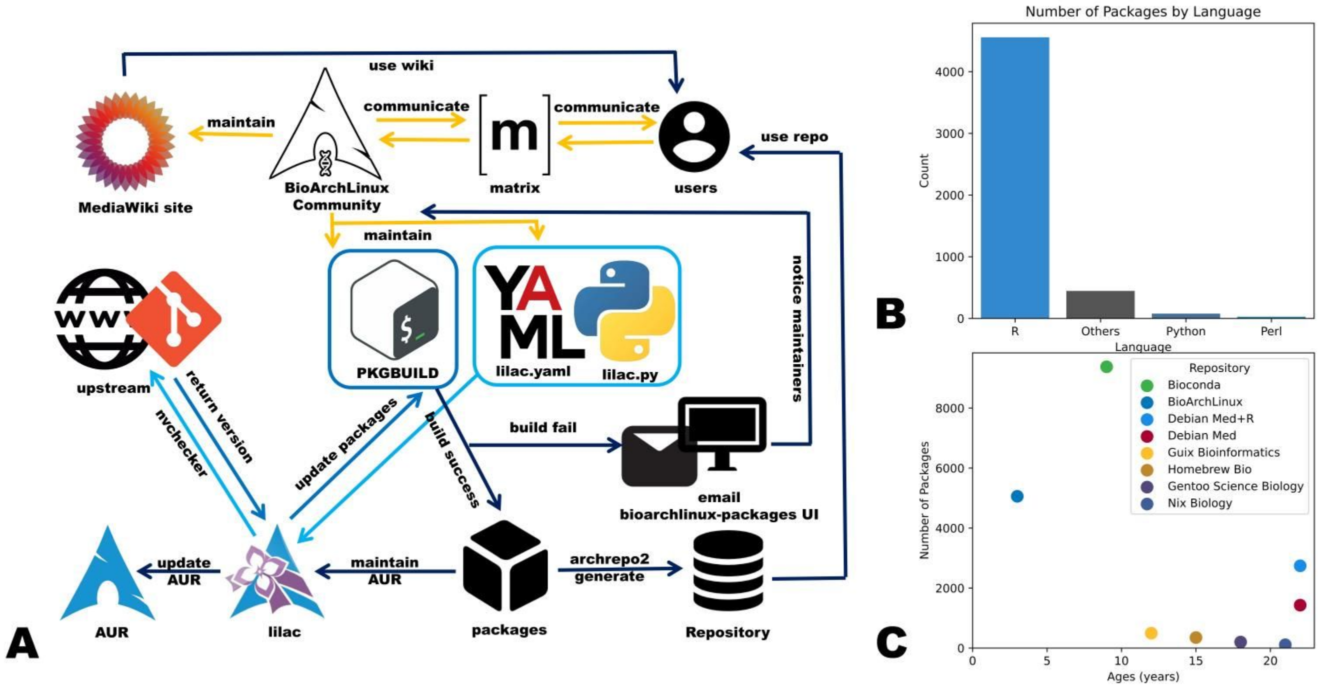

Fig. 1. Workflow and statistics of BioArchLinux. A, Workflow of BioArchLinux community. B, Number of BioArchLinux packages by programming language (Oct 2024). C, Current number of packages and age of different repositories (Oct 2024).

3 Results

Our repositories contain R, Python, Perl and other programming language packages, e.g. C/C++ (Fig. 1B), which breaks the limitation of specific language package managers. We are not the sole initiative distributing bioinformatics software, there are also other repositories available (Fig. 1C). Though we only have 15 maintainers, compared to Bioconda's large team, our repository packages number is more than half of Bioconda's (Fig. 1C). However, this difference is attributed to the distinct development and maintenance philosophies among different Linux distributions

Crucially, our project encourages participation, whether it is for maintaining existing software or releasing new packages. We strive to maintain a high level of freshness and aim to create a comprehensive, language-agnostic, and easy-to-install collection of life sciences software. We welcome everyone to join our community and contribute to the BioArchLinux project. Together, we can create a robust and user-friendly software repository for the life sciences.

Acknowledgements

We extend our sincere thanks to all contributors, users and mirror providers for their support and recognition. We are also grateful to the Open Bioinformatics Foundation for enabling our attendance at the Bioinformatics Open Source Conference (BOSC) free of charge. Special thanks go to xTom (https://xtom.com/) for generously providing a free Linux server for compiling packages and hosting our website. We appreciate the helpful comments from two anonymous reviewers and Frank Köhler (Australian Museum), the consideration of Associate Editor Peter Robinson, and the generosity of Oxford University Press (OUP) for waiving our Article Processing Charge (APC).

Conflict of Interest: none declared.

References

Baker M. 1,500 scientists lift the lid on reproducibility. *Nature* 2016; **533**: 452–454. https://doi.org/10.1038/533452a

Baykal PI, Łabaj PP, Markowetz F *et al*. Genomic reproducibility in the bioinformatics era. *Genome Biol* 2024; **25**: 213. https://doi.org/10.1186/s13059-024-03343-2

Bedő J, Di Stefano L, Papenfuss AT. Unifying package managers, workflow engines, and containers: Computational reproducibility with BioNix. *GigaScience* 2020; **9**: giaa121. https://doi.org/10.1093/gigascience/giaa121

Drexel J, Hänggi E, Veiga IM. Reproducible Builds and Insights from an Independent Verifier for Arch Linux. Bonn, Sicherheit 2024; https://dl.gi.de/handle/20.500.12116/43956

Grüning B, Dale R, Sjödin A *et al*. Bioconda: Sustainable and comprehensive software distribution for the life sciences. *Nat Methods*, 2018; **15**: 475–476. https://doi.org/10.1038/s41592-018-0046-7

Hegde SG, Ranjani G. Package Management System in Linux. 2021 *Asian Conference on Innovation in Technology (ASIANCON)* 2021; 2021:1–6. https://doi.org/10.1109/ASIANCON51346.2021.9544805

Hettrick S. It's impossible to conduct research without software, say 7 out of 10 UK researchers. Software Sustainability Institute. 2014; Retrieved October, 20, 2024.

Horner DS, Pavesi G, Castrignanò T *et al*. Bioinformatics approaches for genomics and post genomics applications of next-generation sequencing. *Brief Bioinform* 2010; **11**: 181–197. https://doi.org/10.1093/bib/bbp046

Hoste K, Timmerman J, Georges A *et al*. EasyBuild: Building Software with Ease. *2012 SC Companion: High Performance Computing, Networking Storage and Analysis* 2012; 572–582. https://doi.org/10.1109/SC.Companion.2012.81

Kanterakis A, Iatraki G, Pityanou K *et al.* Towards Reproducible Bioinformatics: The OpenBio-C Scientific Workflow Environment. *2019 IEEE 19th International Conference on Bioinformatics and Bioengineering (BIBE)* 2019; 221–226. https://doi.org/10.1109/BIBE.2019.00047

Kumar S, Dudley J. Bioinformatics software for biologists in the genomics era. *Bioinformatics* 2007; **23**: 1713–1717. https://doi.org/10.1093/bioinformatics/btm239

Legay D, Decan A, Mens T A Quantitative Assessment of Package Freshness in Linux Distributions. *2021 IEEE/ACM 4th International Workshop on Software Health in Projects, Ecosystems and Communities (SoHeal)* 2021; 9–16. https://doi.org/10.1109/SoHeal52568.2021.00008

Mesirov JP. Accessible Reproducible Research. *Science* 2010; **327**: 415–416. https://doi.org/10.1126/science.1179653

Raghupathi W, Raghupathi V, Ren J. Reproducibility in Computing Research: An Empirical Study. *IEEE Access* 2022; **10**: 29207–29223. https://doi.org/10.1109/ACCESS.2022.3158675

Repology Team. Track and compare package versions in many repositories. 2024; Available at https://repology.org

Squillace F. Junest: The lightweight Arch Linux based distro that runs, without root privileges, on top of any other Linux distro. 2024; Retrieved from https://github.com/fsquillace/junest

Sy H, Irmayana A, Gufran M *et al*. Distribution Linux for Installation Software using Remastering Technique. 2021 *3rd International Conference on Cybernetics and Intelligent System (ICORIS)* 2021; 1–4. https://doi.org/10.1109/ICORIS52787.2021.9649591

Waltemath D, Wolkenhauer O. How Modeling Standards, Software, and Initiatives Support Reproducibility in Systems Biology and Systems Medicine. *IEEE Trans Biomed Eng* 2016; **63**: 1999–2006. https://doi.org/10.1109/TBME.2016.2555481

Yoo C. Open Source, Modular Platforms, and the Challenge of Fragmentation. 2016; Retrieved from https://scholarship.law.upenn.edu/faculty_scholarship/1693