Recommendations to enhance rigor and reproducibility in biomedical research

This manuscript was automatically generated on April 8, 2020.

Authors

- **Jaqueline J. Brito**
  0000-0002-7158-3253 · jaquejbrito · jaquejbrito
  Department of Clinical Pharmacy, School of Pharmacy, University of Southern California

- **Jiu Li**

  Department of Computational Medicine & Bioinformatics, Medical School, University of Michigan

- **Jason H. Moore**

  Department of Biostatistics, Epidemiology, and Informatics, Institute for Biomedical Informatics, University of Pennsylvania

- **Casey S. Greene**
  0000-0001-8713-9213
  Department of Systems Pharmacology and Translational Therapeutics, Perelman School of Medicine, University of Pennsylvania,; Childhood Cancer Data Lab

- **Nicole A. Nogoy**
  0000-0002-5192-9835
  GigaScience

- **Lana X. Garmire**

  Department of Computational Medicine & Bioinformatics, Medical School, University of Michigan

- **Serghei Mangul**
  0000-0003-4770-3443 · smangul1
  Department of Clinical Pharmacy, School of Pharmacy, University of Southern California

Abstract

Computational methods have reshaped the landscape of modern biology. While the biomedical community is increasingly dependent on computational tools, the mechanisms ensuring open data, open software, and reproducibility are variably enforced by academic institutions, funders, and publishers. Publications may present academic software for which essential materials are or become unavailable, such as source code and documentation. Publications that lack such information compromise the role of peer review in evaluating technical strength and scientific contribution. Incomplete ancillary information for an academic software package may bias or limit any subsequent work produced with the tool. We provide eight recommendations across four different domains to improve reproducibility, transparency, and rigor in computational biology — precisely the main values which should be emphasized in life science curricula. Our recommendations for improving software availability, usability, and archival stability aim to foster a sustainable data science ecosystem in biomedicine and life science research.

Keywords: Rigor, reproducible research, installability, archival stability, big data, open science

Introduction

Biomedical informatics is increasingly becoming essential to development of practices that promote open data, open software, and reproducible research in the scientific community. Computational reproduction of previously published results is enabled when scientists publicly release all research resources, from raw data to installable packages and source code, in a discoverable and archivally stable manner. Publications lacking data or source code sharing undermine scientific rigor, transparency, and reproducibility [1]. Platforms already exist that support public release of scientific materials, but the current lack of strict enforcement by journals, academic institutions, and funding agencies has resulted in a loss of essential data and source code for many published studies.

An astonishing number of bioinformatics and computational biology software tools are designed each year to accommodate increasingly bigger, complex, and specialized biomedical datasets [2]. Many of those software tools have limited installability, are closed-source, or are hosted on Uniform Resource Locators (URLs) with undetermined archiving protocols [3]. Lack of access to the source code of a software package undermines the auditing of methods and results and ultimately harms the transparency of research. Prior studies have addressed issues of computational reproducibility, including the need to automatize all data manipulation tasks and version control of code [4]. We expand upon existing dialogue and emphasize reproducible research as computational training, journal policies, and financial support. We identify and discuss eight key recommendations across four different domains (**Figure 1**) to tackle the pressing need for scientists to improve software availability, usability, and archival stability in computational biology. By following a set of best practices [5], scientists can promote rigor and reproducibility, ultimately cultivating a sustainable, thriving research community.

1. Teaching computational skills to produce reproducible research

Increase computational training opportunities targeted at reproducibility.

Biomedical researchers who use computational tools must acquire specific computational skills in order to successfully apply the techniques to a large amount of data. Undergraduate students who lack formal computational training can be taught the skills required to promote reproducibility via specialized courses [6]. In addition to rigorous class training, advanced undergraduate and graduate students, postdoctoral scholars, clinical fellows, and faculty may benefit from short-term intensive workshops [7]. Several institutions, including the University of California, Los Angeles, have successfully

hosted workshop-based programs for over five years and serve as valuable resources for pedagogy and curriculum development 8,9. Effective workshops for training researchers to use computational tools include curated, hands-on training experiences for implementing analysis tools, such as interactive cloud-based notebook technologies. Since 1998, Software Carpentry (https://software-carpentry.org/) has been holding volunteer-based training courses for researchers who wish to master the computational skills required to keep up with the demands of data- and computational-intensive research. Today's biological researcher must learn to use the command line in order to run analyses in open-source software packages. , Comprehensive computational training programs are ideal platforms for training future life science and biomedical researchers in techniques that support reproducibility **(Figure 1a)**.

2. Development and distribution of data and software

2.1. Make all data and metadata open and discoverable

Open source code depends on the availability of open and shareable data, and access to the data used to produce important research results is key for auditing the rigor of published studies. Open access to datasets is imperative to building a thriving and sustainable scientific community where all researchers can access and analyze existing data. In practice, omics data of patients often cannot be publicly shared due to patient privacy and/or user agreement standards 10, 11, 12, 13, 14, 15, 16, 17. While not all data are freely and publicly available, many studies provide controlled data access where researchers can sign a user agreement to access the raw data once their scientific rationale is approved. In general, the global data sharing climate has shifted towards a positive direction; even in cases where raw data are not accessible by the public, summary data are often available.

Truly open data sharing supports the reproducibility and robustness of science because it enables others to reuse data on larger-scale analyses. In addition, secondary analysis is an economically sustainable approach that can be adopted by scientists in countries or at institutions with limited computational resources 18, 19. Ideally, data should also be discoverable via centralized repositories, such as Sequence Read Archive (SRA) and Gene Expression Omnibus (GEO), and annotated with descriptive metadata to enhance data reuse **(Figure 1b)**. When data are shared on centralized repositories in interoperable formats, other researchers can examine and re-analyze the data, challenge existing interpretations, and test new theories. Data sharing corresponds to the true spirit of science, where each new discovery is built upon previous work and ultimately allows us to "stand on the shoulders of giants". Many important scientific discoveries have been solely based on shared data (e.g., economics, meteorology, and physics). Reusing data further emphasizes the quality and importance of generated data and contributes to the impact of the original, data-generating research.

2.2. Build and use open-source software

Software provides a foundation for scientific reproducibility — the ability to replicate published findings by running the same computational tool on data generated by the study 5;6. Open-source academic software are advantageous to the scientific community, because closed-source proprietary software restricts the reproducibility of biomedical research. First, lack of access to the source code limits other researchers' ability to audit results and reviewers' ability to test the reproducibility prior to publication. Second, license restrictions may prohibit the creation of new functionalities that could be released on modified versions of existing tools. Not every laboratory or researcher can afford the cost of acquiring and maintaining proprietary software licenses. Reviewers may lack access to proprietary software and be unable to fully test the reproducibility of results. Widespread adoption of standard open-source licenses for data and software tools can enhance the rigor and impact of research by allowing any researcher and reviewer to reproduce published studies.

Publicly releasing the source code does not guarantee the computational reproducibility of biomedical research. Software must be well documented with user manuals and installable in a user-friendly manner. Code used in a published analysis should be hosted on an archivally stable platform such as Software Heritage Archive (https://archive.softwareheritage.org) or Zenodo (https://zenodo.org) (**Figure 1c**). Currently, over one-fourth of computational software resources cannot be accessed through the URLs provided in the original publication, suggesting that the repositories are poorly maintained [3]. Additionally, many bioinformatics tools are too difficult, or even impossible, for a new user to install [3]. Use of Open Source Initiative license models (https://opensource.org/licenses) allows users to easily use and adapt tools, increasing the sustainability of the biomedical research community. Hosting software tools on package managers allows users to easily install software with more straightforward commands and automatically acquire resolutions for software dependencies. Examples of package managers are Conda and Bioconda **(Table 1).**

We highlight CODE CHECK (https://sje30.github.io/codecheck/) as a promising platform to follow. CODE CHECK allows researchers to input code and data, and check that an analysis is reproducible. For verified analyses, CODE CHECK issues a time-stamped "*certificate of reproducible computation*", which a reviewer of a paper can consider without running the analyses themselves.

2.3. Leverage platforms that enhance reproducibility

In addition to software and datasets, computational biology researchers commonly produce resources such as experiment protocols, workflows, and annotations. Storing and sharing these resources on a stable platform allows other researchers to cite the materials, which would increase the reproducibility of a paper and the visibility of previously developed methods. The inclusion of citable digital object identifiers (DOIs) also facilitates the discovery of reusable resources as they provide long-term access to published resources. Several innovative platforms designed to promote reproducibility have recently emerged (**Figure 1d**).

3. Implementation of reproducible research

3.1. Make tools and workflows reproducible

Virtual machines (VMs) and containers can be used to facilitate the reproducibility of open-source software tools. VMs are software pieces that are capable of encapsulating entire operating systems, libraries, codes, and data. Reproducibility can be enhanced with workflow-specific platforms [20], such as Galaxy and Tensorflow (for machine learning), and workflow standards, such as CWL (Common Workflow Language) **(Table 1)**. Various platforms and tools are now available that support reproducible research and are already commonly used by life science and biomedical researchers (**Table 1**). Given the many different tools and platforms available, a research lab should define their own standards on a suite of tools and platforms that support their research practices (**Figure 1e**).

Table 1: Examples of tools and platforms to share reproducible resources.

Platform & Type	Use
Reproducible and open methods	**Protocols.io** is an open-source protocol repository, where researchers can manage, share, tweak, optimize and adopt existing methods even after a scientist has left a lab.

Platform & Type	Use
RRIDs	**Scicrunch.org** is a platform for curating research resources that enables the user to discover, access, view, and use research objects. Users can register any research object, such as tools, antibodies, animal models etc In turn,. these get issued a RRID (Research Resource ID) which should be cited in the manuscript. The RRID allows other users to easily locate and access the resources.
Annotations	**Hypothes.is** is an open-source annotation tool that allows any researcher to annotate any resource on the web, including to create annotations for personal use or as part of conversations with private groups or the general public.
Virtual Machines & Containers	Containers such as **Docker**(www.docker.com), and **Singularity** (singularity.lbl.gov) are lightweight solutions compared to VMs as they do not encapsulate the operating system; rather, they rely on the host kernel to run required functions. Both VMs and containers are shared via image files and can be included as supplementary material at certain journals or stored in Zenodo (https://zenodo.org/), Figshare (https://figshare.com/), or other general-purpose archival repositories.
Reproducible workflows	**Galaxy** (https://galaxyproject.org/) is a computational platform which allows users to share workflows, histories and wrapped tools in an easy-to-use and open source interface that even people without coding experience can use., **Tensorflow** (https://www.tensorflow.org/) is an open source end-to-end machine learning platform with broad use (e.g., data, library and neural networks). Tensorflow provides workflows to develop and train models using many other programming languages. **Common workflow language (CWL)** (https://www.commonwl.org)- is an open standard used to describe workflows and tools to make them portable and interoperable across different environments e.g. cloud, cluster or high performance computing.
Package managers	**Conda** (https://conda.io/) is a powerful open source package and management system that can quickly install, run and update packages and their dependencies. **Bioconda** (https://bioconda.github.io/) leverages Conda, it is a community project and package manager dedicated to the life sciences and biomedical tools.
Reproducible documents & Figures	**Stencila** (https://stenci.la/) is an open source framework for executable documents and living figures (using R scripts). It supports commonly used environments and tools, such as Juptyer notebook, RMarkdown, Python and SQL. **Jupyter Notebook** (https://jupyter.org/) allows for the creation of sharing of live code, equations, visualizations and narrative text. The application supports over 40 different programming languages and can be used to leverage big data. **MyBinder** (https://mybinder.org/) is an application that collects and 'binds' interactive jupyter notebooks into a Binder repository and can also create a Docker image of that.

3.2. Implement living and reproducible figures and papers

Archiving open data and code is an important fundamental step toward transparency; however, over the last 5 years, it has been possible to break away from the static presentation of results and produce dynamic, or "living," figures (**Figure 1f**). Dynamic figures allow a reader to alter parameters of an analysis as the code is actively running [21] – an iterative process where a data visualization can evolve in real time as new data is added. One such example is Stenci.la, a platform that supports executable documents, living figures, and Jupyter Notebooks (**Table 1**).

4. Incentivizing reproducible research

4.1. Enforce reproducibility upon the peer-reviewing process

Journals have various publishing standards. Stakeholders from academia and industry have defined a set of principles stating that research data should be Findable, Accessible, Interoperable, and Reusable (FAIR) [22] (**Figure 1g**). Researchers may elect to publish in journals that encourage best practices (e.g., adopting the FAIR principles [22] that aim to increase the impact of their work. To ensure reproducibility, many journals now require that biomedical data generated by a published study be shared when the paper is released. For instance, *GigaScience* (https://academic.oup.com/gigascience) has been promoting reproducibility of analyses since 2012 (in addition to publishing open access) by mandating open data arrangements that follow the FAIR principles and mandates availability of source code with an OSI approved license. During peer review, *GigaScience* makes all supporting data and code available for reviewers, and editors ask reviewers to test provided materials for reproducibility. Authors can aid this task by including VMs, containers, Jupyter Notebooks, or packaged workflows (as opposed to static versions of these resources). *Biostatistics* has begun issuing badges for articles with validated data and code sharing [23]. In 2018, *eLife* published their first example of a dynamic and code-based reproducible figure in a peer-reviewed paper, using the Stencila platform (**Table 1**). This approach enables data and analysis to be fully reproducible by the reader [24] and challenges the traditional static representation of results using PDF or HTML formats.

4.2. Create earmarked funds and reporting requirements to support reusable resources

Successfully implementing and widely distributing software tools developed in academia involves unique challenges when compared to doing so in industry. In academia, software tools are developed by small groups comprised of graduate or postdoctoral scholars. These groups have fairly fast turn-over rates of 2-5 years and are less likely to be professionally trained in software production standards [25]. In industry, software development groups are comprised of holistic teams of specialists capable of supporting long-term software maintenance. In order to enhance the quality and reuse of open software, academic groups should hire professionally trained software engineers to partner with students and postdocs. Clearly hiring industry software developers represents a burden on academic teams; funding agencies need clear mechanisms of acknowledging and incentivizing funding earmarked for critical bioinformatics infrastructure (**Figure 1h**). In addition, funders should recognize the rigor of software development, rather than just considering 'novelty'-based conventional criteria of research. The availability of well-resourced grant mechanisms to convert minimum viable products produced by trainees into reliable software could enhance the impact of research-grade software on the community [11]. With the growing number of biomedical datasets open for reuse in the public domain, it is inspiring to see the encouragement and acknowledgment of data reuse and secondary analysis with the Research Parasite Awards [26]. The annual Parasite Awards recognizes the exceptional contributions for rigorous secondary analysis of data with recognition of the top-performing junior parasite and senior parasite. More initiatives, such as this, are needed for promoting software reuse.
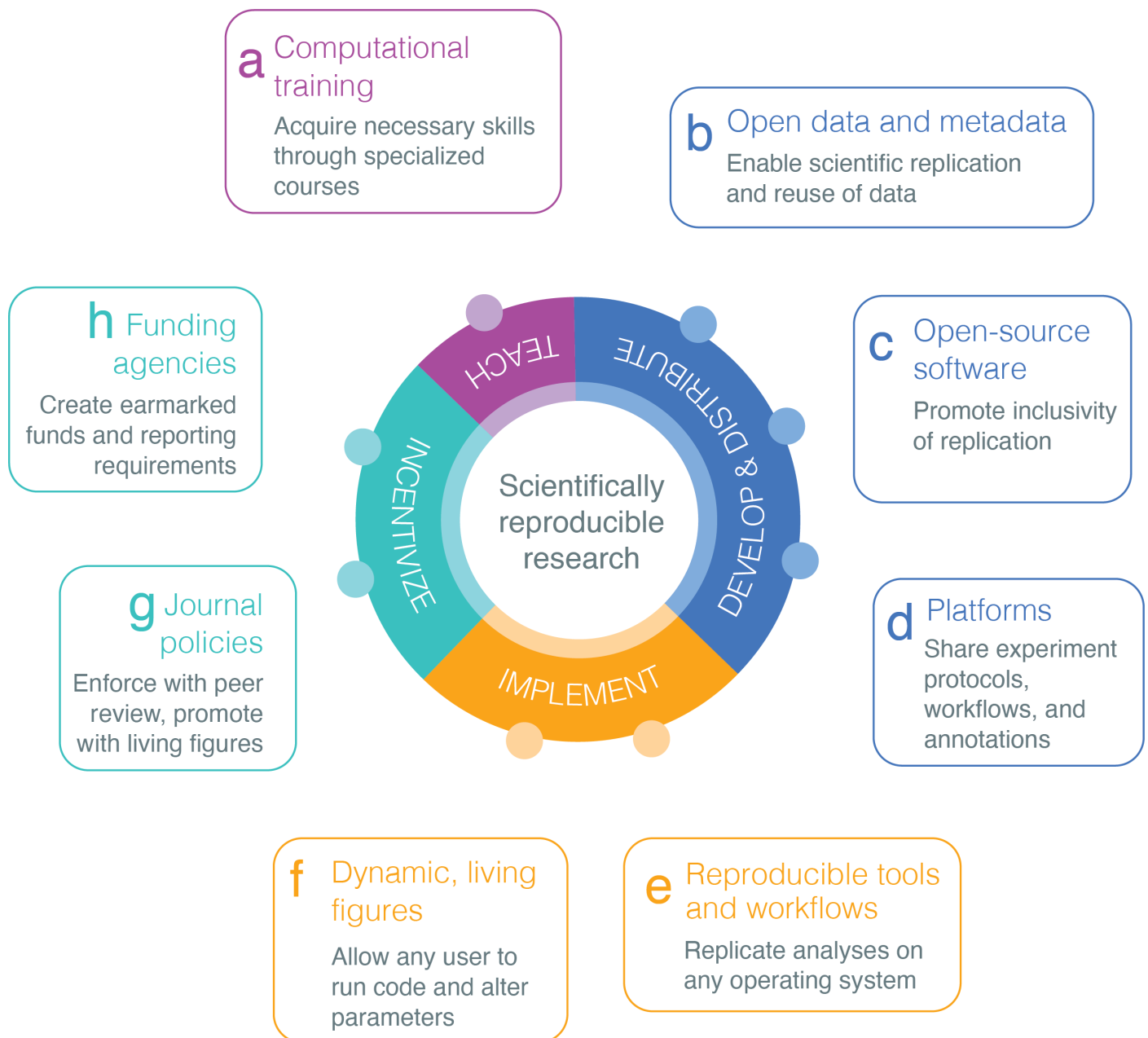
Figure 1. Recommendations to improve reproducibility and rigor of biomedical research organized across the four domains: Teaching computational skills to produce reproducible research ("Teach"); Development and distribution of data and software ("Develop and Distribute"); Implementation of reproducible research ("Implement"); and Incentivizing reproducible research ("Incentivize").

Conclusions

We outlined eight key recommendations across four different domains to improve the rigor of biomedical studies and foster reproducibility in computational biology. The infrastructure required to systematically adopt best practices for reproducibility of biomedical research is largely in place; the remaining challenge is that incentives are not currently aligned to support good practices. Instead, current efforts rely on individual researchers electing to follow the best practices, often at their own time and expense. We believe it is time for a fundamental cultural shift in the scientific community: rigor and reproducibility should become primary concerns in the criteria and decision-making process of designing studies, funding research, and writing and publishing results. Successful systematic adoption of best practices will require the buy-in of multiple stakeholders in the scientific communities, from publishers, academic institutions, funding agencies, and stakeholders. Such commitment would increase the lifetime and scientific value of published research as resources

naturally become reusable, testable, and discoverable. Community-wide adoption of best practices for reproducibility is critical to realizing the full potential of fast-paced, collaborative analyses of large datasets in the biomedical and life sciences. The platforms listed in this paper are provided for illustration. Given this is a fast-moving area, some of our recommendations are likely to be outdated within a short period and other short-lived. We acknowledge that new platforms may appear soon. Therefore we have created this dynamic version of this paper to reflect new platforms, as an example to command scientific rigor and reproducibility ourselves. The dynamic version was compiled in markdown and includes an extended list of references.

Abbreviations

CWL, Common Workflow Language; DOI, Digital Object identifier; FAIR, Findable, Accessible, Interoperable, and Reusable; GEO, Gene Expression Omnibus; OSI, Open Source Initiative; RRID, Research Resource ID; SRA, Sequence Read Archive; URLs, Uniform Resource Locators; VM, Virtual Machine.

Competing interests

The authors declare that they have no competing interests.

Funding

Authors' contributions

Author	Contribution
JJB	Writing - Original Draft Preparation, review and editing.
JL	Writing - review and editing.
CSG	Writing - review and editing.
JM	Writing - review and editing.
NN	Conceptualization and structure of the manuscript; Writing - review and editing; Visualization - creation of Figures and Table.
LG	Conceptualizing the project; Writing - review and editing.
SM	Conceptualization and structure of the manuscript; Writing - review and editing.

Acknowledgments

Authors' Information

NAN is an Editor at *GigaScience* and is an open science advocate with over 7 years experience in publishing reproducible research.

References

1. **An empirical analysis of journal policy effectiveness for computational reproducibility**
   Victoria Stodden, Jennifer Seiler, Zhaokun Ma
   Proceedings of the National Academy of Sciences (2018-03-12) https://doi.org/gc8gkw
   DOI: 10.1073/pnas.1708290115 · PMID: 29531050 · PMCID: PMC5856507

2. **Bioinformatics programs are 31-fold over-represented among the highest impact scientific papers of the past two decades.**
   Jonathan D Wren
   Bioinformatics (Oxford, England) (2016-05-05) https://www.ncbi.nlm.nih.gov/pubmed/27153671
   DOI: 10.1093/bioinformatics/btw284 · PMID: 27153671

3. **Improving the usability and archival stability of bioinformatics software**
   Serghei Mangul, Lana S. Martin, Eleazar Eskin, Ran Blekhman
   Genome Biology (2019-02-27) https://doi.org/ggc6j6
   DOI: 10.1186/s13059-019-1649-8 · PMID: 30813962 · PMCID: PMC6391762

4. **Ten Simple Rules for Reproducible Computational Research**
   Geir Kjetil Sandve, Anton Nekrutenko, James Taylor, Eivind Hovig
   PLoS Computational Biology (2013-10-24) https://doi.org/pjb
   DOI: 10.1371/journal.pcbi.1003285 · PMID: 24204232 · PMCID: PMC3812051

5. **Tools and techniques for computational reproducibility**
   Stephen R. Piccolo, Michael B. Frampton
   GigaScience (2016-07-11) https://doi.org/gfs3cq
   DOI: 10.1186/s13742-016-0135-4 · PMID: 27401684 · PMCID: PMC4940747

6. **Strategies for Building Computing Skills To Support Microbiome Analysis: a Five-Year Perspective from the EDAMAME Workshop**
   Ashley Shade, Taylor K. Dunivin, Jinlyung Choi, Tracy K. Teal, Adina C. Howe
   mSystems (2019-08-20) https://doi.org/ggkt8k
   DOI: 10.1128/msystems.00297-19 · PMID: 31431509 · PMCID: PMC6702294

7. **Making Workshops Work: Insights from EDAMAME**
   Jaclyn N. Taroni
   mSystems (2019-08-20) https://doi.org/ggkt8m
   DOI: 10.1128/msystems.00467-19 · PMID: 31431510 · PMCID: PMC6702295

8. **Addressing the Digital Divide in Contemporary Biology: Lessons from Teaching UNIX.**
   Serghei Mangul, Lana S Martin, Alexander Hoffmann, Matteo Pellegrini, Eleazar Eskin
   Trends in biotechnology (2017-07-15) https://www.ncbi.nlm.nih.gov/pubmed/28720283
   DOI: 10.1016/j.tibtech.2017.06.007 · PMID: 28720283

9. **greenelab/GCB535**
   Greene Laboratory
   (2019-04-19) https://github.com/greenelab/GCB535

10. **Data Access in Genomics** https://www.nature.com/collections/diadgjciaj

11. **Data sharing and the future of science**
    Nature Communications
    (2018-07-19) https://doi.org/cr99
    DOI: 10.1038/s41467-018-05227-z · PMID: 30026584 · PMCID: PMC6053389

12. **No impact without data access**
    Nature Genetics
    (2015-06-26) https://doi.org/ggkt8q
    DOI: 10.1038/ng.3351 · PMID: 26111506

13. **Analysis commons, a team approach to discovery in a big-data environment for genetic epidemiology**
    Jennifer A BrodyAlanna C Morrison, Joshua C Bis, Jeffrey R O'Connell, Michael R Brown, Jennifer E Huffman, Darren C Ames, Andrew Carroll, Matthew P Conomos, …
    Nature Genetics (2017-10-27) https://doi.org/gchz7j
    DOI: 10.1038/ng.3968 · PMID: 29074945 · PMCID: PMC5720686

14. **Analysis of five years of controlled access and data sharing compliance at the International Cancer Genome Consortium** *Nature Genetics* (2016-02-24) https://doi.org/ggkt8r
    DOI: 10.1038/ng.3499 · PMID: 26906679

15. **DNA.Land is a framework to collect genomes and phenomes in the era of abundant genetic information**
    Jie Yuan, Assaf Gordon, Daniel Speyer, Richard Aufrichtig, Dina Zielinski, Joseph Pickrell, Yaniv Erlich
    Nature Genetics (2018-01-26) https://doi.org/gdcc77
    DOI: 10.1038/s41588-017-0021-8 · PMID: 29374253

16. **UK Biobank debuts as a powerful resource for genomic research.**
    Teri A Manolio
    Nature medicine (2018-12) https://www.ncbi.nlm.nih.gov/pubmed/30510254
    DOI: 10.1038/s41591-018-0276-3 · PMID: 30510254

17. **Using and producing publicly available genomic data to accelerate discovery in nephrology**
    Matthew G. Sampson, Hyun Min Kang
    Nature Reviews Nephrology (2019-06-10) https://doi.org/ggkt8s
    DOI: 10.1038/s41581-019-0166-z · PMID: 31182850

18. **Using bioinformatics training to boost research capacities in resource-limited regions**
    Serghei Mangul, Lana Martin, Ben Langmead, Javier Sanchez Galan, Ian Toma, Pavel Pevzner, Eleazar Eskin
    PeerJ (2018-12-06) https://doi.org/ggkt8t
    DOI: 10.7287/peerj.preprints.27415v1

19. **How bioinformatics and open data can boost basic science in countries and universities with limited resources**
    Serghei Mangul, Lana S. Martin, Ben Langmead, Javier E. Sanchez-Galan, Ian Toma, Fereydoun Hormozdiari, Pavel Pevzner, Eleazar Eskin

Nature Biotechnology (2019-03) https://doi.org/gfwnsm
DOI: 10.1038/s41587-019-0053-y · PMID: 30833765

20. **Practical Computational Reproducibility in the Life Sciences**
Björn Grüning, John Chilton, Johannes Köster, Ryan Dale, Nicola Soranzo, Marius van den Beek, Jeremy Goecks, Rolf Backofen, Anton Nekrutenko, James Taylor
Cell Systems (2018-06) https://doi.org/ggdv3z
DOI: 10.1016/j.cels.2018.03.014 · PMID: 29953862 · PMCID: PMC6263957

21. **Referee report. For: Sub-strains of Drosophila Canton-S differ markedly in their locomotor behavior [v1; approved 1, http://f1000r.es/3is]**
Gregg Roman
F1000Research (2014) https://doi.org/ggkt8w
DOI: 10.5256/f1000research.4564.r5630

22. **The FAIR Guiding Principles for scientific data management and stewardship**
Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, … Barend Mons
Scientific Data (2016-03-15) https://doi.org/bdd4
DOI: 10.1038/sdata.2016.18 · PMID: 26978244 · PMCID: PMC4792175

23. **Badges for sharing data and code at Biostatistics: an observational study**
Anisa Rowhani-Farid, Adrian G. Barnett
F1000Research (2018-03-07) https://doi.org/gdsnhw
DOI: 10.12688/f1000research.13477.2 · PMID: 29862016 · PMCID: PMC5843843

24. **Replication Study: Transcriptional amplification in tumor cells with elevated c-Myc**
L Michelle Lewis, Meredith C Edwards, Zachary R Meyers, C Conover Talbot Jr, Haiping Hao, David Blum, Elizabeth Iorns, Rachel Tsui, Alexandria Denis, Nicole Perfito, …
eLife (2018-01-09) https://doi.org/gf9h4w
DOI: 10.7554/elife.30274 · PMID: 29313490 · PMCID: PMC5760205

25. **The anatomy of successful computational biology software**
Stephen Altschul, Barry Demchak, Richard Durbin, Robert Gentleman, Martin Krzywinski, Heng Li, Anton Nekrutenko, James Robinson, Wayne Rasband, James Taylor, Cole Trapnell
Nature Biotechnology (2013-10) https://doi.org/gf8nr4
DOI: 10.1038/nbt.2721 · PMID: 24104757 · PMCID: PMC4166496

26. **Celebrating parasites.**
Casey S Greene, Lana X Garmire, Jack A Gilbert, Marylyn D Ritchie, Lawrence E Hunter
Nature genetics (2017-03-30) https://www.ncbi.nlm.nih.gov/pubmed/28358134
DOI: 10.1038/ng.3830 · PMID: 28358134 · PMCID: PMC5710834