 doi:10.20944/preprints202206.0320.v1

Article

Ten Simple Rules for Using Public Data for Your Research

Vishal H. Oza[1], Jordan H. Whitlock[1], Elizabeth J. Wilk[1], Angelina Uno-Antonison[2,3,4], Brandon Wilk[2,3,4], Manavalan Gajapathy[2,3,4], Timothy C. Howton[1], Austyn Trull[2,3,4], Lara Ianov[5], Elizabeth A. Worthey[2,3,4] and Brittany N. Lasseigne[1*]

1. Department of Cell, Developmental and Integrative Biology, Heersink School of Medicine, The University of Alabama at Birmingham, Birmingham, AL
2. Center for Computational Genomics and Data Sciences, Heersink School of Medicine, The University of Alabama at Birmingham, Birmingham, AL
3. Department of Pediatrics, Heersink School of Medicine, The University of Alabama at Birmingham, Birmingham, AL
4. Department of Pathology, Heersink School of Medicine, The University of Alabama at Birmingham, Birmingham, AL
5. Civitan International Research Center, Heersink School of Medicine, The University of Alabama at Birmingham, Birmingham, AL

* Correspondence: Author: bnp0001@uab.edu

Abstract: With an increasing amount of "omics" data available publicly, there is a need for a guide on how to successfully download and use this data. The 10 simple rules for using public data are: 1) use public data in your research, 2) evaluate data for your use case, 3) check data reuse requirements and embargoes, 4) be aware of ethics for data reuse, 5) plan for data storage and compute requirements, 6) know what you are downloading, 7) download programmatically and verify integrity, 8) properly cite data, 9) make data FAIR and share, and 10) make pipelines and code FAIR and share. These rules are intended as a guide for researchers wanting to make use of available data and to increase data reuse and reproducibility.

Keywords: data; reproducibility; FAIR; data reuse; public data; big data; analysis

1. Introduction

In recent years, with the advent of high-throughput sequencing technologies, advances in microscopy, and the growth of single-cell technologies, biology is set to overtake other data-heavy disciplines such as astronomy in terms of data storage needs [1]. There has been a dramatic increase in the number and size of data sets deposited by individual labs on data storage servers like the Gene Expression Omnibus (GEO) and dbGAP [2,3] and made available by large consortium efforts such as The Cancer Genome Atlas (TCGA) [4], the Genotype-Tissue Expression (GTEx) project [5], Bgee [6], Human Cell Atlas [7], ENCODE [8,9], etc. Additionally, the commitment by funders, publishers, and individual scientists to make data sets (especially those funded by taxpayers and donors) publicly available has rapidly increased opportunities for data reuse by the broader scientific community. From January 2023, NIH will now require researchers to share data generated with NIH funds under the new Data Management and Sharing Policy [10] further leading to an increase in the availability of public data. With this growing data deluge, it seemed timely to provide guidelines on why, when, and how investigators can incorporate these valuable resources into their research programs. Here we discuss 10 Simple Rules for using public data with the intention it will serve as a useful guide (Figure 1). While this article focuses on computational biology and bioinformatics, the principles outlined here generally apply to other domains as well.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 June 2022 doi:10.20944/preprints202206.0320.v1

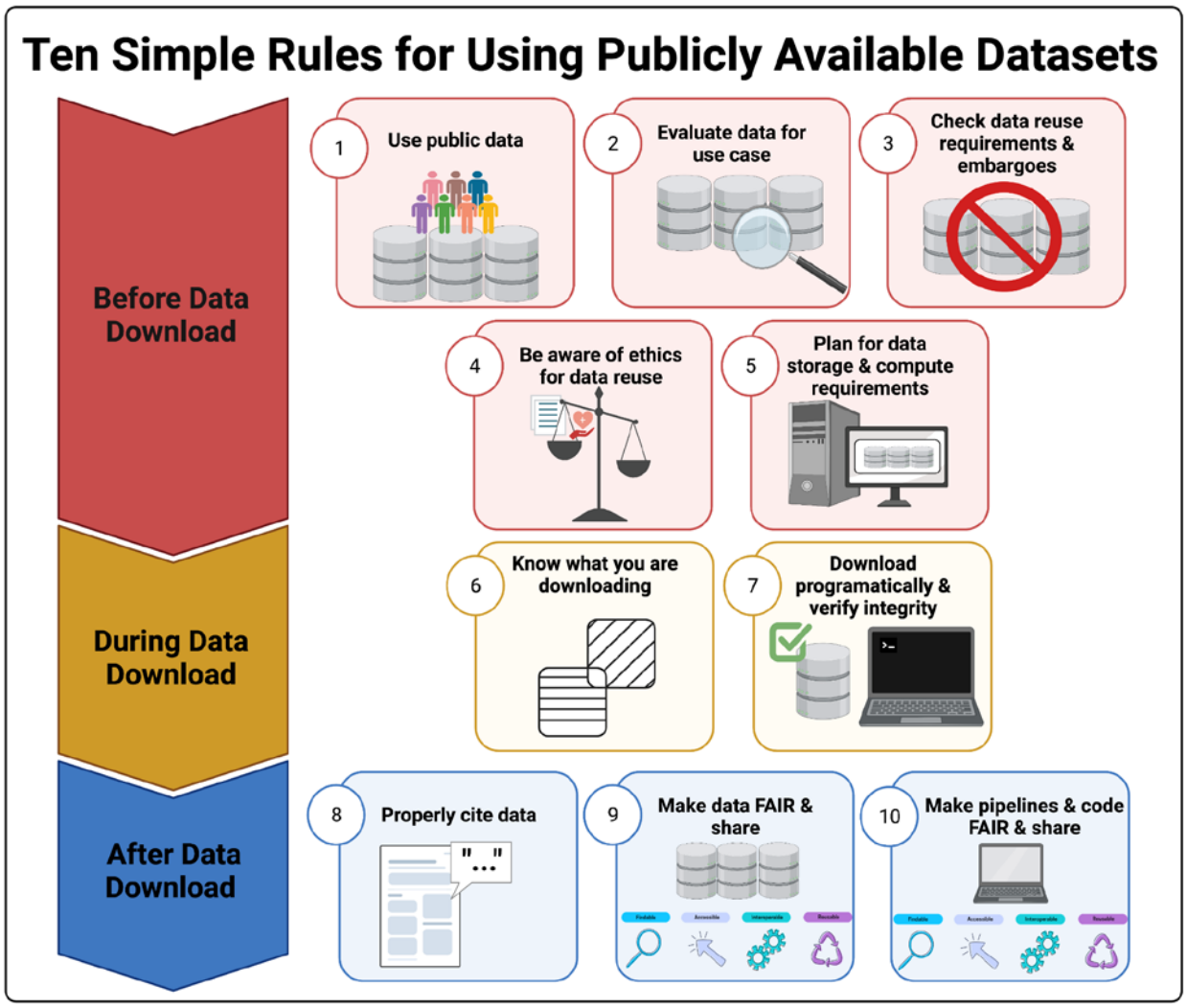

Figure 1. 10 These ten simple rules for using public data span checklist items for pre, during, and post data download.

2. Use public data in your research

Reasons for reusing public research data include cost-effectiveness and efficiency, access to datasets that would be difficult or impossible to regenerate, an increased sense of community, greater transparency and clarity of research, ability to retest and validate a shared dataset, support for recognition of data ownership, and an increase in citations [11]. The use of public data provides individuals with a unique opportunity to leverage existing data sets in combination with other public data or their own. This allows for additional contexts to be explored within your own research. Comparing your research findings with a public data set analyzed in the same way can give nuance or clarity to results, especially in either further confirming what was observed (i.e., validation) or producing different results that help you to re-evaluate your approach or its applicability across cohorts. For example, you can validate current findings in public data, expand the number of samples, age groups, or other parameters of your current analysis, or explore molecular changes in a different system. Assessing public data can also aid in placing your research into the context of, and in perspective to, current research with public data, therefore, allowing comparisons to other work in the field.

Incorporating public data in your research project can also allow you to ask novel questions that the data generator may not have originally foreseen, further moving your own research and the field in new directions. Outside of expanding your research through exploration of new contexts, you can also use public data sets to generate hypotheses and/or guide future research by refining hypotheses with public data for preliminary

 doi:10.20944/preprints202206.0320.v1

analyses. With computational biology becoming ubiquitous in life science research, you can also use public data to drive novel method development and build modeling frameworks for systems biology. In fact, there have been efforts to standardize data sets for ease of use in such analysis [12]. In addition to improving your analysis, using public data also provides an opportunity to practice reproducible and interoperable research and further develop professional data science skills [11].

In short, novel studies do not always require new data, and research including data reuse not only offers all of the benefits mentioned here but can support community building and greater impact; as shown by Milham et al [13] who found that neuroimaging research making use of public data was cited at the same or higher frequency in higher impact journals as compared to research using only self-generated data (this is expanded upon more in Rule 8).

3. Evaluate data carefully for your use case

Before we get down to the business of downloading the data, we need to make sure we get the right goods. As scientists, we are interested in finding signals and/or patterns in our data. However, patterns in a data set can arise from many different variables. Therefore, it is essential to carefully evaluate the available data for your particular use case, otherwise, the results obtained could be wrong and/or misleading. It is critical to examine the metadata (the data describing the data set) before downloading the actual data. Some important metadata variables to consider (but by no means exhaustive) relate to 1) the nature of the sample (e.g., the specific animal model or cell line(s)), genotype, age, sex, 2) sample collection data (e.g., the type of sample, collected, time points of collection, 3) the platforms used (e.g., the sequencing or imaging platform), and 4) data quality metrics. Not considering these variables can dramatically influence downstream analysis, interpretation, and therefore the selection of data set to use.

Once the data set is deemed appropriate for your analytical goals, it is key to check for the confounding effects of metadata variables that might impact the statistical model and cause you to reach wrong conclusions. This includes a review of data distributions and sampling as well as performing exploratory analysis and visualizations such as principal component analysis (PCA) and correlation analysis. “10 simple rules for initial data analysis”[14] provides excellent pointers which can be applied to metadata analysis as well.

4. Be aware of, and adhere to, data reuse requirements, embargoes, etc.

Researchers looking to reuse controlled-access data (elaborated more in Rule 4) assume the responsibility (along with their legal and financial administrative offices) to protect the rights and welfare of the original participants. As such, data repositories with controlled-access data sets develop Data Access Request (DAR) processes to appropriately restrict access. DAR processes usually include agreeing to terms of access for the requested data, a Data Use Certification Agreement (DUA) between the requester’s institution and the repository, a description of the intended use, and an acknowledgment agreement. If a data set is available in a data portal (e.g., GEO [2], SRA [15], etc.) that does not require controlled access, it is considered open-access and can be freely downloaded through the portal. Even under these circumstances, the original data generators may still request acknowledgment upon use (a practice that should be followed whether or not the generators have requested it). A third possibility for acquiring controlled-access data for reuse is to contact the corresponding author of the paper directly to request data. However, this method can prove difficult for many reasons related to the author’s availability to fulfill the request (e.g., protected time to fulfill the request, ability to transfer the data, willingness to comply, etc.). Therefore, independent of the method of requesting access to restricted data, it is important to be patient, but persistent. Recently, many journals

 doi:10.20944/preprints202206.0320.v1

have begun implementing data availability manuscript sections that outline the portal where the data is stored and can be accessed.

Regardless of how the data is acquired, it is important to be aware of the legal, regulatory, and security obligations associated with its use. For example, data licenses might be in place to protect the original data generator's rights by permitting secondary parties to reuse the data according to specific restrictions. Because countries have different regulations for data reuse [16], data licenses may clarify the uncertainty of requirements for data reusers. It is important to understand by which license or waiver the data to be reused is governed. In an effort to promote open access to data, many journals and data repositories operate under Creative Commons (CC BY 4.0) licenses that allow the freedom to share and adapt data as long as the original data generators are acknowledged for their contribution. For example, PLOS ONE stipulates that if the data associated with a published article in their journal is deposited in a repository with a licensing agreement, the agreement can not be more restrictive than CC BY [17]. Researchers should identify which, if any, data license is governing the data they wish to reuse and respect any limitations associated with it.

Some data generators place an embargo on the data they generate in order to ensure they have time to publish initial findings. The data set may be submitted to a public repository but unavailable for download or publication for a certain length of time or until a specified date. Additionally, with the increased prevalence of pre-printing articles, data generators may wish to withhold their data sets until their article is published in a peer-reviewed journal. Being aware of data access and publication restrictions associated with the data set of interest and understanding who sets data restrictions can be complicated (e.g., funders, consortia, journals, individual labs, etc.), but convention, or in some cases the legal requirement, is to follow the stated restrictions.

5. Be aware of ethical considerations like confidentiality and protected health information

The privacy and ethics of data sharing are critical and it is the responsibility of researchers to protect and observe [18]. With regards to data reuse, you must verify that the data you are planning to reuse was ethically collected or generated, was collected for a purpose in alignment with additional applications, and ensure that the study does not use the data irresponsibly or immorally. It is our duty as scientists and citizens to move science forward while protecting sensitive data and presenting studies fairly. Special considerations need to be made for data reuse:

Ethics by data type: The ethics of sharing and using data depend heavily on the data type. For example, cell lines, animal models, and microbial data are typically very low risk for privacy or ethical issues; that is not to say, however, that they have no risk (e.g., sharing pathogenic sequence data) or alluding to locations for endangered species that could increase the risk of poaching [19]. Ethical concerns are most common when working with human data.

HIPAA, PHI, and patient de-identification: The Health Insurance Portability and Accountability Act (HIPAA) of 1996 Privacy Rule is a federal law regarding protected health information (PHI) for individuals and their access to their own health information, as well as the specific permissible use and disclosure of PHI with other organizations [20]. Any purposeful or accidental disclosure of PHI as a HIPAA violation can lead to hefty fines. Health data used in research needs to be de-identified in a manner that makes re-identification highly unlikely unless that is the willful goal of the project. Genomics data is particularly susceptible to re-identification due to the uniquely identifying nature of the data itself; particularly when coupled with geographic collection metadata such as collection site. Well-constructed data usage agreements and controlled-access repositories can mitigate re-identification risks (see Rule 3).

 doi:10.20944/preprints202206.0320.v1

Consent: Consent is an agreement between healthcare groups and participants that adheres to The Privacy Rule and HIPAA authorizations. For example, the GTEx Live Donor Informed Consent Template (BBRB-PM-0018 [21]) asserts exclusion of access to participant PHI, that the generated data will be saved for many years, that it will be available to scientists around the world, and that data maybe used broadly for medical research. It is important to ensure that data reuse does not violate any initially obtained consent.

Ethics specific to how/where data were obtained: Data may be obtained directly from an individual lab, or queried from a private or public repository. Publicly-available data that can be downloaded by anyone tends to be the least risk for violating the privacy, and are the easiest to access (i.e. TCGA [4] and GTEx [5] gene expression data). However, when data is received from another investigator through direct sharing or from a controlled-access repository, it becomes your obligation to secure that data and not further share it with unauthorized individuals [19].

Ethical design for data reuse: When reusing data, ensure fairness and equality with your representation of the data, including but not limited to ancestry or sex. For example, genetic study participation has been disproportionately overrepresented by European descendants, where one study found that as of 2018, individuals in GWAS catalog consisted of 78% European, 10% Asian, 2% African, 1% Hispanic, and <1% for all other races [22]. In GTEx as of 2022, 84.6% of donors were of White origin, 12.9% of African American origin, 1.3% of Asian origin, 1.1% unknown, and no statistic for Hispanic/Latino origin [5]. Additionally, sex differences impact every area of health and have been largely disregarded in study design and have also been heavily unbalanced; especially in pharmaceutical trials where women were previously excluded entirely due to potential pregnancies during trials [23]. Since then, women are now included in study designs, though sex specificity has not been accounted for leading to vastly more adverse drug reactions in women due to inappropriate drug and dose recommendations [24]. Careful consideration of these factors will lead to more rigorous and accurate results and avoid perpetuating these issues in research.

6. Plan for needed data storage and compute requirements

Ask yourself, is my data "genomical" in size? [25] It is if you're working in genomics, but regardless of your data, the amount of data being integrated into research is growing dramatically and the cost for storage and computation is at a rate not seen previously. Knowledge is power, and in this case, it is knowledge of the resources that will be needed *before* you need it that will benefit your work most. Data storage and computing hardware requirements should be determined and documented prior to downloading any data sets. This avoids potentially time-consuming and expensive surprises by 1) identifying gaps in the current infrastructure, and 2) allowing expert support staff an opportunity to investigate the viable solutions, where possible, in a timely manner. While needs vary by the individual situation and domain, here we discuss various criteria to determine your needs. Data size, type, level of access, and security are the major considerations for where data needs to be stored. All invested parties who will retrieve, process, and consume the data need to be identified and involved in answering the following questions:

Data size:

- What is the estimated size of data to be retrieved?
- How many individual files are included?
- How much data is expected to be produced during processing?
- Do you need backups of the original or processed data?

Data type:

- Are the data types large or small?
- Are the file types binary or textual?
- Are the files compressed?
- Access requirements:
- Who needs (or does not need) access to the data?

 doi:10.20944/preprints202206.0320.v1

- What is the level of access required?
- How often will access be needed?
- How often will reanalysis be necessary?
- Do any of the users require (or prefer) a data-sharing platform with a graphic interface for ease of access?
- Are there any institutional policies or approvals to be considered prior to granting access?

Data security [26]

- Does the data include PHI?
- Does the access need to be restricted?
- What is the minimum level of data security required?
- How will they be secured?
- Will the data need to be deleted after a certain time period or event? Why, when, and how?
- Are there any institutional policies or approvals to be considered?
- Who will supervise data security?
- How often(e.g., half-yearly, annually, etc.) will the adopted policies and implementations need to be reviewed and verified?

Answers to these questions will narrow down the data storage options(s) and dictate if other institutional entities such as IT or the office of sponsored research need to be involved. Further, costs associated with data storage and reanalysis needs should also be considered when selecting storage locations. Commonly used data storage locations are personal computer(s), HPC environments, and commercial cloud storage services (e.g., Dropbox, Box, AWS, etc.). Depending on the data, it may need to be split based on the type and stored in a data-type-specific location best suited for its purpose. For example, large data such as FASTQ and VCF files may need to be stored at a location where they are accessible from an HPC environment for downstream processing, whereas spreadsheets and text documents may need to be stored in a cloud storage service (e.g., Box) to facilitate accessibility for non-computational team members.

Computing or hardware requirements will depend on the type of data and the processing planned [27]. A personal computer might suffice to process small data sets, but access to HPC or cloud computing (e.g., AWS, Microsoft Azure, Google Cloud, etc.) is often needed to process large data sets. HPCs and cloud computing offer several advantages such as fast processors, multicore chips, higher memory resources, and Graphics Processing Units (GPU), which enable parallelization and scalability of a large number of jobs.

7. Know what you are downloading

As outlined in Rule 2, downstream analysis can be affected by how the data was collected and processed. This becomes an even greater problem when trying to integrate multiple public data sets, each collected and processed separately. So, be mindful of the processing differences between them. In recent years, a number of projects have been established that use and share common data processing pipelines across data collections, allowing researchers to more easily combine data for more complex analysis. For example, recount3 [28] is a uniformly processed resource of hundreds of thousands of human and mouse RNA-Seq data sets designed to facilitate meta-analysis and cross-study comparisons. Another great resource is the Bgee [6] database which contains “normal”, healthy, wild-type expression data across 52 different species thereby providing a comparable reference of gene expression by anatomical entities within and between species. A third example is BioDataome [29]. BioDataome has ~5,600 human and mouse expression and methylation data sets pre-processed in an analogous manner to allow for direct comparisons between data. All of these data repositories provide R packages facilitating programmatic data download (see Rule 7). Collections of equivalently processed data are becoming more of the norm, but datasets of interest may not be included in them. When this is

 doi:10.20944/preprints202206.0320.v1

the case, it may be necessary to reprocess the data to minimize the downstream impact of differences in upstream processing.

8. Download data programmatically; verify data integrity

Data downloads should be performed in a programmatic manner for consistency, scalability, and reliability. Several tools have been developed to provide direct programmatic access to large public databases (e.g., SRA [15], GEO [2], ENA [30], TCGA [4]) and they should be implemented when possible. One example of these tools is the SRA toolkit [31] which contains a number of commands linked to data download (e.g., `fasterq-dump` may be implemented to download FASTQs and `vdb-validate` to check the integrity of SRA data sets). Additional examples include database-specific computational packages/libraries such as Bioconductor packages `recount3`[28] and `GenomicDataCommons`[32]. A key step in the data download process is validating the checksum (typically the MD5 hash) provided in the hosting database to verify data integrity prior to data analysis and interpretation. While these checks can be performed manually, some tools such as the Genomic Data Commons (GDC) Data Transfer Tool Client [32], will automatically validate MD5 checksums as part of the download process.

Depending on the number of files and tool functionality and design, customized scripts or workflows (e.g., Snakemake [33], NextFlow [34]) may be needed for scalability. When possible, publicly available workflows that adopt workflow management systems are recommended. One example is `fetchngs` [35] provided by nf-core [36]. This Nextflow pipeline [34] uses tools such as sra-tools [31] to download FASTQ files and metadata based on a user-provided accession ID list. At the time of writing, fetchngs [35] supports the use of SRA [15], ENA [30], DDBJ [37], GEO [2], and Synapse [38] IDs. When performing downloads in this way, documenting the sample identifier and database sources becomes a critical step for reproducibility. For studies deposited in databases such as GEO [2], recording the GEO accession number along with sample identifiers may be sufficient, however, other data sources are updated on a continuous basis (e.g., Genotype-Tissue Expression (GTEx) Portal [5]). In such cases, the database version (e.g., GTEx Analysis Release V8 - dbGaP Accession phs000424.v8.p2) and the date the download is performed should be recorded and reported in publications.

These practices should also be applied to data types beyond sample data including, for example, genome references (e.g., genome/transcriptome FASTA files, annotations files, etc.) from Ensembl [39], GENCODE [40], UCSC [41], NCBI [42], and others. Checksums from genomic reference files should be validated and details linked to the files should be recorded, including the FASTA file type (e.g., primary assembly or entire genomic sequence which includes assembly patches and haplotypes), database name, assembly name, and version or release number. The same principles should be applied to other reference files such as GTF/GFF [43].

9. Be a good community member; properly cite data

As with any journal article or publication, researchers who generated the data you used must be credited and properly cited. This practice benefits the original data generator by providing a tangible demonstration of value and impact beyond the initial data publication. Failing to cite the data source withholds credit from researchers in the same way that failing to cite a journal article does. Public data citations support better quality and more transparent science, making a compelling argument for other researchers to contribute their own data to public data repositories. This process also supports improved reproducibility and credibility for your own research.

When it comes to citing data, the field still lacks a gold standard. However, existing best practices include citing the original paper where the data was published or a data object identifier (doi) generated from a persistent database like Zenodo [44] or figshare

 doi:10.20944/preprints202206.0320.v1

[45]. In addition, consortium projects such as GTEx [5] and TCGA [4] often provide guidance on citation practices for their repositories. In an effort to increase the citability of public data and establish the data itself as a scientific output of value separate from the associated manuscript, many researchers now publish data sets independently from their associated publications. Journals for this purpose such as *Scientific Data* now exist [46]. If there is no guideline associated with the data set in general the citation should include the generators, where the data was obtained from, accession numbers, the version number for the data (if applicable), and the date it was accessed.

Non-profit organizations such as DataCite and Crossref provide unique persistent identifiers (PIDs) to data sets to improve tracking usage and facilitate linking to the publication of origin [47]. Check and see if the public data you are using in your analysis has a PID that can be cited or included in the methods section. In the future, PIDs may perhaps be linked to an individual's ORCID number to provide a standardized data citation approach. A study looking at the correlation between if data was publicly available and the citation rate of the original paper demonstrated that those including publicly available data within the paper were associated with a 69% increase in citations [48]. Furthermore, investigators who share public data sets will have an increase in the impact of their own research. Articles with links to data repositories or that include PIDs are more highly cited than those without [19]. In summary, public data use benefits both the creator and the user.

10. Make data FAIR and share

All of the rules mentioned above are possible to adhere to because researchers made their data Findable, Accessible, Interoperable, and Reusable (FAIR) [49]. As a contributing member of the research ecosystem, you should pass it on, too! Make sure the data you generate, as well as reuse, adheres to the FAIR principles. If your research is funded through the NIH then you must adhere to a Data Management and Sharing Plan as outlined in the NIH policy starting in 2023 [10]. Recent research shows that only 6.8% of authors respond to requests for data sharing dramatically reducing the impact of most data and the knowledge to be gained from its use [50]. To be FAIR:

Improve Data Findability: Submit your data to stable open-access public data repositories that provide DOIs such as Zenodo [44] and figshare [45]. Personal, lab, group, or institute sites are not good long-term solutions. Socialize your data by sharing it on social media platforms such as Twitter. Blogs and news articles on lab and/or institute sites, and presenting at conferences where data are clearly identified with DOIs helps others to discover useful data. By diversifying where and how you share with the community about your data, you cast the widest net in order to catch the attention of potential researchers who would also benefit from access to your data.

Improve Data Accessibility: Open access to publications is important to science and plays a critical role in reproducing and advancing science. Making your data readily accessible is equally important. Share your raw as well as processed data so that the analysis performed can be reproduced fully and with minimal effort.

Improve Data Interoperability: Interoperability refers to the ability of data from different sources to be able to integrate with minimal effort [49]. A good example is the FHIR standard for health care data exchange [51]. This becomes even more critical in studies where data from different sources are being reused, so make sure the data you provide is highly interoperable by including relevant metadata and adhering to appropriate and reasonable file and data conventions within the field.

Improve Data Reuse: Make your data reuse, reusable. Providing data in standard and popular formats goes a long way in making it reusable. Incomplete metadata and methods can severely limit the reuse (and usefulness) of data. Don't skimp by providing the bare minimum data and metadata needed to satisfy the requirements of the granting body, governing body, journal, etc., that you're looking to communicate your work through. Abstaining from providing all necessary assets to reproduce the work does a disservice to

 doi:10.20944/preprints202206.0320.v1

you, your colleagues, your lab and institute, and the scientific community. It is not just about the input and the output: there's a whole bunch of research, development, refinement, knowledge, and other work done in between data input and output that needs to be captured, codified, and shared with the work itself.

11. Make your processing, pipelines, and code FAIR

Tools and methods used for the analysis and interpretation of public data sets should also adhere to the FAIR guidelines for coding and software development [52]. Many of the FAIR principles for data (outlined in rule 9) are directly applicable to software, but others require modification for application to software [53]. For instance, persistent identifiers should be generated and recorded for novel pipelines, software, and research tools. Findability of software should be linked to the traceability of the source code under version control (e.g., GitHub [54], GitLab [55], BitBucket [56]), and reporting of appropriate metadata such as software versions. Similarly, software and code associated with analysis should be accessible through repositories and (when appropriate) software archives such as CRAN [57], Bioconductor [58], PyPI [59], and Conda [60]. The same is true for software dependencies. Software containers (e.g., Docker [61], Singularity [62]) can also be implemented to enable software portability and analysis reproducibility. In the absence of containerized software, the necessary information for how to build and install a published tool should be provided.

Because the dynamic nature of software can mean the most up-to-date version is different from the version that was published, proper documentation and tagging (e.g., GitHub tags linked to released versions of software, etc.) allow researchers to find the exact package versions used during a study, therefore facilitating reproducible analyses. Continuous integration approaches (the software development practice of automating builds, static analysis, and tests on code changes that were pushed to a central repository) can automate time-consuming tasks associated with pipeline and software development. Similarly, static code analysis tools in continuous integration pipelines can help automate source code quality analysis allowing early identification of potential bugs, security vulnerabilities, performance issues, or deviations from the project/organization and coding guidelines.

The inclusion of continuous integration and static code analysis tools allows for rapid feedback loops in code inspection, reducing the time reviewers need to spend reviewing and reducing the cost of time and funding of development maintenance. Journals are increasingly requiring that the code used for analysis is made available [63]. In addition to ensuring that your processing, pipelines, and code are FAIR, the steps above will help to ensure code review is not an onerous task being performed during the submission process and your methods are reproducible for other scientists. For more detailed information about how you can make your research more computationally reproducible refer to [64].

12. Conclusion

In summary, data reuse is not only good for science, it is the right thing to do in order to extract the greatest societal impact from the samples and funding that patients, donors, and taxpayers generously provide. Here we covered 10 Simple Rules for data reuse spanning the periods before, during, and after data download. This paper serves as a guide for both data users and generators in the community.

CRediT Author Statement: Vishal H. Oza: Conceptualization, Writing - Original Draft, Writing - Review & Editing, Project administration. **Jordan H. Whitlock:** Writing - Original Draft, Visualization, Writing - Review & Editing, Funding acquisition (T32GM008111). **Elizabeth J. Wilk:** Writing - Original Draft, Writing - Review & Editing. **Angelina Uno-Antonison:** Writing - Original Draft. **Brandon Wilk:** Writing - Original Draft, Writing - Review & Editing. **Manavalan Gajapathy:** Writing - Original Draft, Writing - Review & Editing. **Timothy C. Howton:** Writing - Original Draft, Writing - Review & Editing. **Austyn Trull:** Writing - Original Draft, Writing - Review & Editing.

Lara Ianov: Writing - Original Draft, Writing - Review & Editing, Supervision, Funding acquisition (UAB CIRC funds)**. Elizabeth A. Worthey:** Writing - Review & Editing, Supervision, Funding acquisition (U54OD030167, UAB Worthey Lab funds, UG1HD107688, WORTHE19A0). **Brittany N. Lasseigne:** Conceptualization, Writing - Original Draft, Writing - Review & Editing, Supervision, Project administration, Funding acquisition (U54OD030167, R00HG009678, R03OD030604, UAB Lasseigne Lab Startup funds).

Funding and Acknowledgements: VHO, EJW, AUA, LI, and EAW, and BNL are supported by U54OD030167. VHO, TCH, and BNL are supported by R00HG009678. VHO and BNL are supported by R03OD030604. JHW is supported by T32GM008111. JHW, TCH, BNL are supported by UAB Lasseigne Lab Startup funds. AUO, BW, MG, AT, and EAW are supported by UAB Worthey Lab funds. BW and EAW are supported by UG1HD107688. BW, MG, and EAW are supported by the CF Foundation (WORTHE19A0). LI is supported by UAB CIRC funds. AT, LI, EAW, and BNL are members of the UAB Biological Data Science (U-BDS) Core (RRID:SCR_021766).

References

1. Navarro FCP, Mohsen H, Yan C, Li S, Gu M, Meyerson W, et al. Genomics and data science: an application within an umbrella. Genome Biol. 2019;20: 109.

2. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, et al. NCBI GEO: archive for functional genomics data sets--update. Nucleic Acids Res. 2013;41: D991–5.

3. Mailman MD, Feolo M, Jin Y, Kimura M, Tryka K, Bagoutdinov R, et al. The NCBI dbGaP database of genotypes and phenotypes. Nat Genet. 2007;39: 1181–1186.

4. The Cancer Genome Atlas Program. In: National Cancer Institute [Internet]. 13 Jun 2018 [cited 6 Jun 2022]. Available: https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga

5. GTEx Portal. [cited 26 Apr 2022]. Available: https://gtexportal.org/home/tissueSummaryPage

6. Bastian FB, Roux J, Niknejad A, Comte A, Fonseca Costa SS, de Farias TM, et al. The Bgee suite: integrated curated expression atlas and comparative transcriptomics in animals. Nucleic Acids Res. 2021;49: D831–D847.

7. Regev A, Teichmann SA, Lander ES, Amit I, Benoist C, Birney E, et al. The Human Cell Atlas. Elife. 2017;6. doi:10.7554/eLife.27041

8. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489: 57–74.

9. Luo Y, Hitz BC, Gabdank I, Hilton JA, Kagda MS, Lam B, et al. New developments on the Encyclopedia of DNA Elements (ENCODE) data portal. Nucleic Acids Res. 2020;48: D882–D889.

10. Kozlov M. NIH issues a seismic mandate: share data publicly. Nature. 2022;602: 558–559.

11. Boté J-J, Termens M. Reusing Data Technical and Ethical Challenges. DESIDOC Journal of Library & Information Technology. 2019. pp. 329–337. doi:10.14429/djlit.39.06.14807

12. Introduction. In: Alevin-fry requant [Internet]. [cited 24 May 2022]. Available: https://combine-lab.github.io/quantaf/

13. Milham MP, Craddock RC, Son JJ, Fleischmann M, Clucas J, Xu H, et al. Assessment of the impact of

shared brain imaging data on the scientific literature. Nat Commun. 2018;9: 2818.

14. Baillie M, le Cessie S, Schmidt CO, Lusa L, Huebner M, Topic Group "Initial Data Analysis" of the STRATOS Initiative. Ten simple rules for initial data analysis. PLoS Comput Biol. 2022;18: e1009819.

15. Leinonen R, Sugawara H, Shumway M, International Nucleotide Sequence Database Collaboration. The sequence read archive. Nucleic Acids Res. 2011;39: D19–21.

16. Labastida I, Margoni T. Licensing FAIR data for reuse. Data Intellegence. 2020;2: 199–207.

17. PLOS ONE. [cited 13 Jun 2022]. Available: https://journals.plos.org/plosone/s/data-availability

18. Bonomi L, Huang Y, Ohno-Machado L. Privacy challenges and research opportunities for genomic data sharing. Nat Genet. 2020;52: 646–654.

19. Byrd JB, Greene AC, Prasad DV, Jiang X, Greene CS. Responsible, practical genomic data sharing that accelerates research. Nat Rev Genet. 2020;21: 615–629.

20. Office for Civil Rights (OCR). Summary of the HIPAA Privacy Rule. [cited 26 Apr 2022]. Available: https://www.hhs.gov/hipaa/for-professionals/privacy/laws-regulations/index.html

21. GTEx Informed Consent Template. [cited 26 Apr 2022]. Available: https://biospecimens.cancer.gov/resources/sops/library.asp

22. Sirugo G, Williams SM, Tishkoff SA. The Missing Diversity in Human Genetic Studies. Cell. 2019;177: 26–31.

23. Oertelt-Prigione S, Mariman E. The impact of sex differences on genomic research. Int J Biochem Cell Biol. 2020;124: 105774.

24. Zucker I, Prendergast BJ. Sex differences in pharmacokinetics predict adverse drug reactions in women. Biol Sex Differ. 2020;11: 32.

25. Stephens ZD, Lee SY, Faghri F, Campbell RH, Zhai C, Efron MJ, et al. Big Data: Astronomical or Genomical? PLoS Biol. 2015;13: e1002195.

26. Hart EM, Barmby P, LeBauer D, Michonneau F, Mount S, Mulrooney P, et al. Ten Simple Rules for Digital Data Storage. PLoS Comput Biol. 2016;12: e1005097.

27. Brandies PA, Hogg CJ. Ten simple rules for getting started with command-line bioinformatics. PLoS Comput Biol. 2021;17: e1008645.

28. Wilks C, Zheng SC, Chen FY, Charles R, Solomon B, Ling JP, et al. recount3: summaries and queries for large-scale RNA-seq expression and splicing. bioRxiv. bioRxiv; 2021. doi:10.1101/2021.05.21.445138

29. Lakiotaki K, Vorniotakis N, Tsagris M, Georgakopoulos G, Tsamardinos I. BioDataome: a collection of uniformly preprocessed and automatically annotated datasets for data-driven biology. Database . 2018;2018. doi:10.1093/database/bay011

30. EMBL-EBI. European Nucleotide Archive. [cited 6 Jun 2022]. Available: https://www.ebi.ac.uk/ena/browser/home

31. Sequence Read Archive Toolkit. [cited 6 Jun 2022]. Available: https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=software

32. Morgan MT, Davis SR. *GenomicDataCommons:* a Bioconductor Interface to the NCI Genomic Data Commons. doi:10.1101/117200

33. Mölder F, Jablonski KP, Letcher B, Hall MB, Tomkins-Tinch CH, Sochat V, et al. Sustainable data analysis with Snakemake. F1000Res. 2021;10: 33.

34. Di Tommaso P, Chatzou M, Floden EW, Barja PP, Palumbo E, Notredame C. Nextflow enables reproducible computational workflows. Nat Biotechnol. 2017;35: 316–319.

35. Patel H, Beber ME, Han DW, Ewels P, Espinosa-Carrasco J, Bot N-C, et al. nf-core/fetchngs: nf-core/fetchngs v1.5 - Copper Cat. 2021. doi:10.5281/zenodo.5746702

36. Ewels PA, Peltzer A, Fillinger S, Patel H, Alneberg J, Wilm A, et al. The nf-core framework for community-curated bioinformatics pipelines. Nat Biotechnol. 2020;38: 276–278.

37. Fukuda A, Kodama Y, Mashima J, Fujisawa T, Ogasawara O. DDBJ update: streamlining submission and access of human data. Nucleic Acids Res. 2021;49: D71–D75.

38. Bionetworks S. Synapse. [cited 6 Jun 2022]. Available: https://www.synapse.org/

39. Ensembl. [cited 6 Jun 2022]. Available: https://www.ensembl.org/index.html

40. Frankish A, Diekhans M, Ferreira A-M, Johnson R, Jungreis I, Loveland J, et al. GENCODE reference annotation for the human and mouse genomes. Nucleic Acids Res. 2019;47: D766–D773.

41. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, et al. The human genome browser at UCSC. Genome Res. 2002;12: 996–1006.

42. National Center for Biotechnology Information. [cited 6 Jun 2022]. Available: https://www.ncbi.nlm.nih.gov/

43. GFF3 - GMOD. [cited 6 Jun 2022]. Available: http://gmod.org/wiki/GFF3

44. European Organization for Nuclear Research, OpenAIRE. Zenodo. CERN; 2013. doi:10.25495/7GXK-RD71

45. Figshare. [cited 6 Jun 2022]. Available: https://figshare.com/

46. van den Berghe GJS-ASV, editor. Scientific Data. Nature Publishing Group; 2014-Current.

47. Pierce HH, Dev A, Statham E, Bierer BE. Credit data generators for data reuse. In: Nature Publishing Group UK [Internet]. 4 Jun 2019 [cited 6 Jun 2022]. doi:10.1038/d41586-019-01715-4

48. Piwowar HA, Day RS, Fridsma DB. Sharing detailed research data is associated with increased citation rate. PLoS One. 2007;2: e308.

49. Wilkinson MD, Dumontier M, Aalbersberg IJJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016;3: 160018.

50. Gabelica M, Bojčić R, Puljak L. Many researchers were not compliant with their published data sharing statement: mixed-methods study. J Clin Epidemiol. 2022. doi:10.1016/j.jclinepi.2022.05.019

51. Index - FHIR v4.3.0. [cited 10 Jun 2022]. Available: http://hl7.org/fhir/index.html

52. Lamprecht A-L, Garcia L, Kuzak M, Martinez C, Arcila R, Martin Del Pico E, et al. Towards FAIR principles for research software. Data sci. 2020;3: 37–59.

53. Jiménez RC, Kuzak M, Alhamdoosh M, Barker M, Batut B, Borg M, et al. Four simple recommendations to encourage best practices in research software. F1000Res. 2017;6. doi:10.12688/f1000research.11407.1

54. Github. In: Github [Internet]. [cited 6 Jun 2022]. Available: https://github.com/

55. Gitlab. In: Gitlab [Internet]. [cited 6 Jun 2022]. Available: https://about.gitlab.com/

56. Bitbucket. In: Bitbucket [Internet]. [cited 6 Jun 2022]. Available: https://bitbucket.org/product/

57. The Comprehensive R Archive Network. [cited 6 Jun 2022]. Available: https://cran.r-project.org/

58. Bioconductor - Home. [cited 6 Jun 2022]. Available: https://bioconductor.org/

59. PyPI · The Python Package Index. In: PyPI [Internet]. [cited 6 Jun 2022]. Available: https://pypi.org/

60. Conda — Conda documentation. [cited 6 Jun 2022]. Available: https://docs.conda.io/en/latest/

61. Docker. [cited 6 Jun 2022]. Available: https://www.docker.com/

62. SingularityCE. In: Sylabs [Internet]. 31 Mar 2022 [cited 6 Jun 2022]. Available: https://sylabs.io/singularity/

63. Cadwallader L, Mac Gabhann F, Papin J, Pitzer VE. Advancing code sharing in the computational biology community. PLoS Comput Biol. 2022;18: e1010193.

64. Heil BJ, Hoffman MM, Markowetz F, Lee S-I, Greene CS, Hicks SC. Reproducibility standards for machine learning in the life sciences. Nat Methods. 2021;18: 1132–1135.