Reproducibility and Transparency by Design

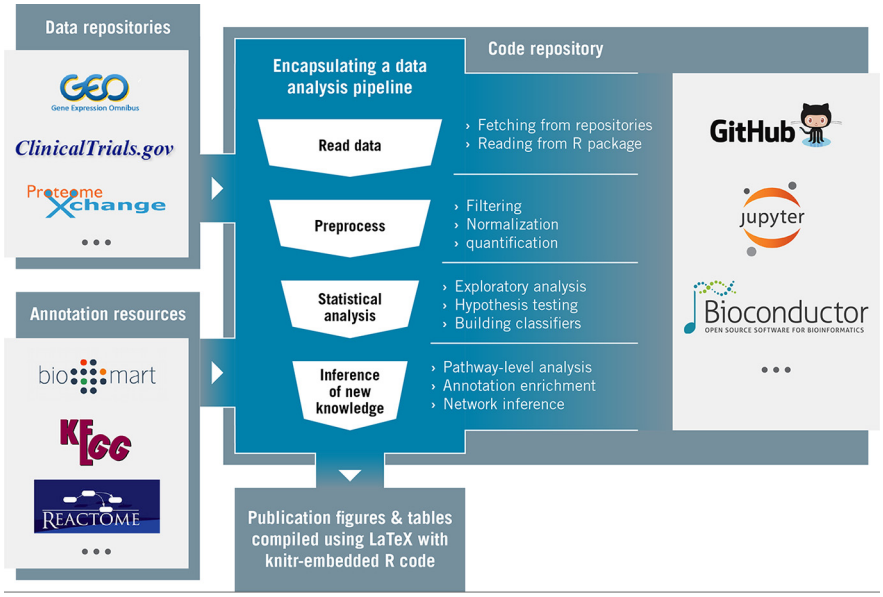
Authors

Vladislav A. Petyuk, Laurent Gatto, and Samuel H. Payne

Correspondence

sam_payne@byu.edu

Graphical Abstract



In Brief

The reproducibility of bioinformatics analyses can be elevated to equal status with biological discovery. To achieve this, reproducibility must become part of the process, not an afterthought.

Highlights

- Repositories are enabling sharing of data.
- Sharing analyses remains a major stumbling block for reproducibility.
- Three simple steps can enable sharing analyses, with nominal effort.

Petyuk et al., 2019, Molecular & Cellular Proteomics *18*, S202–S204

https://doi.org/10.1074/mcp.IP119.001567

⌘ *Author's Choice*

Reproducibility and Transparency by Design*

Vladislav A. Petyuk‡, Laurent Gatto§, and Samuel H. Payne¶‖

To truly achieve reproducible research, having reproducible analytics must be a principal research goal. Biological discovery is not the only deliverable; reproducibility is an essential part of our research.

Public trust of scientific research is affected by the clarity of published conclusions and also the perceived transparency of the method. Although irreproducibility is not exclusive to biology, strong public interest in environmental and biomedical discoveries seems to have focused the spotlight here following a number of high-profile studies that failed to be reproduced (1–6). In this report, we specifically focus on the linked issues of reproducibility and transparency of integration and analyses for multi-omics data. Unlike data generation where biological variability is expected to be manifest, computational analyses should be completely and exactly reproducible. Unfortunately, the documentation of data processing, analysis, and statistical algorithms in publications is usually not sufficiently detailed. This lack of detail is especially problematic for multi-omics characterizations where the complex statistical integration is essential to merging disparate data types (*e.g.* clinical, proteomics, genomics, etc.).

Making Reproducibility a Priority. Where Are the Gaps?—There are many stages of a multi-omics project, and recent efforts have made significant improvement on transparency of data files and selected steps of analysis. *MCP* and other journals have been leaders in requiring the complete sharing of raw data and preliminary processing (7–9). In a multi-omics project, it is now common to require that the mass spectrometry instrument files are freely shared via public repositories, which exist for genomics (10), proteomics (11), and metabolomics (12). Spectral identification must also be reported with the software and associated parameters. Pipelines run through the BioContainers (13) facilitate this recording. Although some popular tools with a graphical interface may not currently store this workflow meta-data, we feel that this is rapidly becoming a demand of both the users and publishers.

After obtaining quantitative molecular data, there is still a lot of work before publication. This includes merging genomic and proteomic data tables, binning samples into phenotypic groups based on multi-omic clustering, functional enrichment analysis, metabolic network modeling, and so on. Unfortunately, the current efforts to mandate data sharing have focus on just the data. Data interpretation and statistical analyses that support scientific conclusions are an equally essential component of our work and must also be openly shared. We write this commentary to highlight the need for greater efforts in the open sharing of analyses.

Although it is a narrow topic, we feel it is important to discuss. As mandated data sharing resolves a portion of the overall transparency/reproducibility challenge, the unaddressed issue remains the sharing of analyses. Moreover, our solution is not that difficult to implement for the new generation of data savvy researchers. It does not require large grants to fund computational/storage infrastructure; it can be done by individual researchers with a modicum of effort. Thus, without delay, journals can start to encourage or enforce the open sharing of computational and statistical data interpretation.

As its central feature, our solution encapsulates the entire data analysis in software, including the creation of publication quality figures. We want to make it easy for peers to do exactly the same analyses in a publication—specifically the critical final steps where data interpretation happens. For example, when discussing an assertion in the results section, it is common to parenthetically list the *p* value and a specific test. To increase the transparency and reproducibility of this assertion, we should share the actual software code that produced this *p* value. Multiple modern software platforms have made this level of transparency achievable with modest effort, including Jupyter notebooks and R markdown (14, 15). Our support for these technologies is not meant to be exclusive but merely convenient as many publications already utilize Python and R/Bioconductor (16). We strongly advocate for the following three steps: code for analysis and figures posted to an open version control software repository like GitHub (17), data tables used by the analysis be posted in the same repository or linked to a password-free download if too large, and the URL to specific scripts in a repository be prominently listed in figure legends and methods sections. The effect of these three would be that anyone interested in a specific figure or conclusion of the paper could easily find the

From the ‡Pacific Northwest National Laboratory, Richland, WA; §de Duve Institute, Université Catholique de Louvain, Brussels, Belgium; ¶Brigham Young University, Provo UT

Received May 14, 2019, and in revised form, June 24, 2019
Published, MCP Papers in Press, July 4, 2019, DOI 10.1074/mcp.IP119.001567

exact analysis method and fully repeat the computation. Indeed, this approach for reproducibility has already been used in a few exemplary publications (18–21).

Looking Forward—The benefits of true transparency have been previously noted (22, 23), and we reiterate that our proposed solution has lasting positive effects for the principal investigator, funding agencies, peer review, collaborators, and the general public. The solution is flexible and applicable to the broad needs of multi-omics integration for climate research, clinical proteogenomics, systems biology, computational neuroscience, and so on.

As multi-omics measurements continue to revolutionize environmental and biomedical research, biology more explicitly becomes a data science. Most graduate programs now require statistics courses, where students learn tools like R and Python. Given the enormous societal impact that comes from scientific discoveries, the transparency of our data and methodology is a critical component of the scientific venture. As large data repositories have begun to capture much of the raw data generated for experiments, we have suggested a companion method to disseminate and expose data analysis methods. Ultimately, the transparency of full disclosure will expose any actual problems underlying irreproducibility in a manner where other researchers can help to correct and advance science.

* V.A.P. was supported by NIH/NINDS U18NS082140. L.G. was supported by BBSRC Strategic Longer and Larger grant (Award BB/L002817/1). S.H.P. was supported by NIH/NCI U24 CA210972 and the U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research, Early Career Research Program. Battelle operates PNNL for the DOE under contract DE-AC05-76RLO01830. The authors declare no competing financial interest.

‖ To whom correspondence should be addressed. E-mail: sam_payne@byu.edu.

Author contributions: V.A.P., L.G., and S.H.P. designed research; V.A.P., L.G., and S.H.P. performed research; V.A.P., L.G., and S.H.P. analyzed data; and V.A.P., L.G., and S.H.P. wrote the paper.

REFERENCES

1. Reaves, M. L., Sinha, S., Rabinowitz, J. D., Kruglyak, L., and Redfield, R. J. (2012) Absence of detectable arsenate in DNA from arsenate-grown GFAJ-1 cells. *Science* **337,** 470–473
2. Petricoin, E. F., Ardekani, A. M., Hitt, B. A., Levine, P. J., Fusaro, V. A., Steinberg, S. M., Mills, G. B., Simone, C., Fishman, D. A., Kohn, E. C., and Liotta, L. A. (2002) Use of proteomic patterns in serum to identify ovarian cancer. *Lancet* **359,** 572–577
3. Kern, S. E. (2012) Why your new cancer biomarker may never work: Recurrent patterns and remarkable diversity in biomarker failures. *Cancer Res.* **72,** 6097–6101
4. Diamandis, E. P. (2010) Cancer biomarkers: Can we turn recent failures into success? *J. Natl. Cancer Inst.* **102,** 1462–1467
5. Li, M., Wang, I. X., Li, Y., Bruzel, A., Richards, A. L., Toung, J. M., and Cheung, V. G. (2011) Widespread RNA and DNA sequence differences in the human transcriptome. *Science* **333,** 53–58
6. Asara, J. M., Schweitzer, M. H., Freimark, L. M., Phillips, M., and Cantley, L. C. (2007) Protein sequences from mastodon and *Tyrannosaurus rex* revealed by mass spectrometry. *Science* **316,** 280–285
7. Kinsinger, C. R., Apffel, J., Baker, M., Bian, X., Borchers, C. H., Bradshaw, R., Brusniak, M. Y., Chan, D. W., Deutsch, E. W., Domon, B., Gorman, J., Grimm, R., Hancock, W., Hermjakob, H., Horn, D., Hunter, C., Kolar, P., Kraus, H. J., Langen, H., Linding, R., Moritz, R. L., Omenn, G. S., Orlando, R., Pandey, A., Ping, P., Rahbar, A., Rivers, R., Seymour, S. L., Simpson, R. J., Slotta, D., Smith, R. D., Stein, S. E., Tabb, D. L., Tagle, D., Yates, J. R. 3rd, and Rodriguez, H. (2011) Recommendations for mass spectrometry data quality metrics for open access data (corollary to the Amsterdam Principles). *Mol. Cell. Proteomics* **10,** O111.015446
8. Chalkley, R. J., MacCoss, M. J., Jaffe, J. D., and Röst, H. L. (2019) Initial Guidelines for Manuscripts employing data-independent acquisition mass spectrometry for proteomic analysis. *Mol. Cell. Proteomics* **18,** 1–2
9. Abbatiello, S., Ackermann, B. L., Borchers, C., Bradshaw, R. A., Carr, S. A., Chalkley, R., Choi, M., Deutsch, E., Domon, B., Hoofnagle, A. N., Keshishian, H., Kuhn, E., Liebler, D. C., MacCoss, M., MacLean, B., Mani, D. R., Neubert, H., Smith, D., Vitek, O., and Zimmerman, L. (2017) New guidelines for publication of manuscripts describing development and application of targeted mass spectrometry measurements of peptides and proteins. *Mol. Cell Proteomics* **16,** 327–328
10. Grossman, R. L., Heath, A. P., Ferretti, V., Varmus, H. E., Lowy, D. R., Kibbe, W. A., and Staudt, L. M. (2016) Toward a shared vision for cancer genomic data. *New Eng. J. Med.* **375,** 1109–1112
11. Deutsch, E. W., Csordas, A., Sun, Z., Jarnuczak, A., Perez-Riverol, Y., Ternent, T., Campbell, D. S., Bernal-Llinares, M., Okuda, S., Kawano, S., Moritz, R. L., Carver, J. J., Wang, M., Ishihama, Y., Bandeira, N., Hermjakob, H., and Vizcaíno, J. A. (2017) The ProteomeXchange consortium in 2017: Supporting the cultural change in proteomics public data deposition. *Nucleic Acids Res.* **45,** D1100–D1106
12. Wang, M., Carver, J. J., Phelan, V. V., Sanchez, L. M., Garg, N., Peng, Y., Nguyen, D. D., Watrous, J., Kapono, C. A., Luzzatto-Knaan, T., Porto, C., Bouslimani, A., Melnik, A. V., Meehan, M. J., Liu, W. T., Crusemann, M., Boudreau, P. D., Esquenazi, E., Sandoval-Calderon, M., Kersten, R. D., Pace, L. A., Quinn, R. A., Duncan, K. R., Hsu, C. C., Floros, D. J., Gavilan, R. G., Kleigrewe, K., Northen, T., Dutton, R. J., Parrot, D., Carlson, E. E., Aigle, B., Michelsen, C. F., Jelsbak, L., Sohlenkamp, C., Pevzner, P., Edlund, A., McLean, J., Piel, J., Murphy, B. T., Gerwick, L., Liaw, C. C., Yang, Y. L., Humpf, H. U., Maansson, M., Keyzers, R. A., Sims, A. C., Johnson, A. R., Sidebottom, A. M., Sedio, B. E., Klitgaard, A., Larson, C. B., Torres-Mendoza PCABD, Gonzalez, D. J., Silva, D. B., Marques, L. M., Demarque, D. P., Pociute, E., O'Neill, E. C., Briand, E., Helfrich, E. J. N., Granatosky, E. A., Glukhov, E., Ryffel, F., Houson, H., Mohimani, H., Kharbush, J. J., Zeng, Y., Vorholt, J. A., Kurita, K. L., Charusanti, P., McPhail, K. L., Nielsen, K. F., Vuong, L., Elfeki, M., Traxler, M. F., Engene, N., Koyama, N., Vining, O. B., Baric, R., Silva, R. R., Mascuch, S. J., Tomasi, S., Jenkins, S., Macherla, V., Hoffman, T., Agarwal, V., Williams, P. G., Dai, J., Neupane, R., Gurr, J., Rodriguez, A. M. C., Lamsa, A., Zhang, C., Dorrestein, K., Duggan, B. M., Almaliti, J., Allard, P. M., Phapale, P., Nothias, L. F., Alexandrov, T., Litaudon, M., Wolfender, J. L., Kyle, J. E., Metz, T. O., Peryea, T., Nguyen, D. T., VanLeer, D., Shinn, P., Jadhav, A., Muller, R., Waters, K. M., Shi, W., Liu, X., Zhang, L., Knight, R., Jensen, P. R., Palsson, B. O., Pogliano, K., Linington, R. G., Gutierrez, M., Lopes, N. P., Gerwick, W. H., Moore, B. S., Dorrestein, P. C., and Bandeira, N. (2016) Sharing and community curation of mass spectrometry data with global natural products social molecular networking. *Nature Biotechnol.* **34,** 828–837
13. da Veiga Leprevost, F., Gruning, B. A., Alves Aflitos, S., Röst, H. L., Uszkoreit, J., Barsnes, H., Vaudel, M., Moreno, P., Gatto, L., Weber, J., Bai, M., Jimenez, R. C., Sachsenberg, T., Pfeuffer, J., Vera Alvarez, R., Griss, J., Nesvizhskii, A. I., and Perez-Riverol, Y. (2017) BioContainers: An open-source and community-driven framework for software standardization. *Bioinformatics* **33,** 2580–2582
14. Lamport, L. (1994) *Latex*, Addison-Wesley
15. Baumer, B., Cetinkaya-Rundel, M., Bray, A., Loi, L., and Horton, N. J. (2014) R Markdown: Integrating a reproducible analysis tool into introductory statistics. *arXiv preprint arXiv:*1402.1894
16. Huber, W., Carey, V. J., Gentleman, R., Anders, S., Carlson, M., Carvalho, B. S., Bravo, H. C., Davis, S., Gatto, L., Girke, T., Gottardo, R., Hahne, F., Hansen, K. D., Irizarry, R. A., Lawrence, M., Love, M. I., MacDonald, J., Obenchain, V., Oleś, A. K., Pagès, H., Reyes, A., Shannon, P., Smyth, G. K., Tenenbaum, D., Waldron, L., and Morgan, M. (2015) Orchestrating high-throughput genomic analysis with Bioconductor. *Nat. Methods* **12,** 115–121
17. Perez-Riverol, Y., Gatto, L., Wang, R., Sachsenberg, T., Uszkoreit, J., Leprevost Fda, V., Fufezan, C., Ternent, T., Eglen, S. J., Katz, D. S., Pollard, T. J., Konovalov, A., Flight, R. M., Blin, K., and Vizcaíno, J. A.

(2016) Ten simple rules for taking advantage of Git and GitHub. *PLoS Comput. Biol.* **12,** e1004947

18. Ohnishi, Y., Huber, W., Tsumura, A., Kang, M., Xenopoulos, P., Kurimoto, K., Oleś, A. K., Arauzo-Bravo, M. J., Saitou, M., Hadjantonakis, A. K., and Hiiragi, T. (2014) Cell-to-cell expression variability followed by signal reinforcement progressively segregates early mouse lineages. *Nat. Cell Biol.* **16,** 27–37
19. Laufer, C., Fischer, B., Billmann, M., Huber, W., and Boutros, M. (2013) Mapping genetic interactions in human cancer cells with RNAi and multiparametric phenotyping. *Nat Methods* **10,** 427–431
20. Eglen, S. J., Weeks, M., Jessop, M., Simonotto, J., Jackson, T., and Sernagor, E. (2014) A data repository and analysis framework for spontaneous neural activity recordings in developing retina. *GigaScience* **3,** 3
21. Lee, J. Y., Fujimoto, G. M., Wilson, R., Wiley, H. S., and Payne, S. H. (2018) Blazing Signature Filter: S library for fast pairwise similarity comparisons. *BMC Bioinform.* **19,** 221
22. Markowetz, F. (2015) Five selfish reasons to work reproducibly. *Genome Biol.* **16,** 274
23. McKiernan, E. C., Bourne, P. E., Brown, C. T., Buck, S., Kenall, A., Lin, J., McDougall, D., Nosek, B. A., Ram, K., Soderberg, C. K., Spies, J. R., Thaney, K., Updegrove, A., Woo, K. H., and Yarkoni, T. (2016) How open science helps researchers succeed. *eLife* **5,** e16800