AMERICAN SOCIETY FOR MICROBIOLOGY | Microbiology Resource Announcements

EDITORIAL

Future-Proofing Your *Microbiology Resource Announcements* Genome Assembly for Reproducibility and Clarity

David A. Baltrus,[a] Christina A. Cuomo,[b] John J. Dennehy,[c,d] Julie C. Dunning Hotopp,[e,f,g] Julia A. Maresca,[h] Irene L. G. Newton,[i] David A. Rasko,[e,f] Antonis Rokas,[j,k,l] Simon Roux,[m] Jason E. Stajich[n]

[a]School of Plant Sciences, University of Arizona, Tucson, Arizona, USA
[b]Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA
[c]Queens College of The City University of New York, Queens, New York, USA
[d]The Graduate Center of The City University of New York, New York, New York, USA
[e]Institute for Genome Sciences, University of Maryland Baltimore, Baltimore, Maryland, USA
[f]Department of Microbiology and Immunology, University of Maryland Baltimore, Baltimore, Maryland, USA
[g]Greenebaum Comprehensive Cancer Center, University of Maryland Baltimore, Baltimore, Maryland, USA
[h]Department of Civil and Environmental Engineering, University of Delaware, Newark, Delaware, USA
[i]Department of Biology, Indiana University, Bloomington, Indiana, USA
[j]Department of Biological Sciences, Vanderbilt University, Nashville, Tennessee, USA
[k]Department of Biomedical Informatics, Vanderbilt University, Nashville, Tennessee, USA
[l]Vanderbilt Genetics Institute, Vanderbilt University Medical Center, Nashville, Tennessee, USA
[m]DOE Joint Genome Institute, Walnut Creek, California, USA
[n]Department of Microbiology and Plant Pathology and Institute for Integrative Genome Biology, University of California—Riverside, Riverside, California, USA

ABSTRACT Descriptions of resources, like the genome assemblies reported in *Microbiology Resource Announcements*, are often frozen at their time of publication, yet they will need to be interpreted in the midst of continually evolving technologies. It is therefore important to ensure that researchers accessing published resources have access to all of the information required to repeat, interpret, and extend these original analyses. Here, we provide a set of suggestions to help make certain that published resources remain useful and repeatable for the foreseeable future.

There are many ways to sequence and assemble a genome, with the number of available sequencing and assembly platforms seemingly growing every week. Within sequencing platforms, library preparation, chemistry, and error profiles frequently change. Our primary goal as *Microbiology Resource Announcements* (MRA) editors is to ensure that a manuscript's techniques and protocols are thoroughly documented so that readers can understand the strengths and weaknesses not only of a particular genome assembly but also the underlying raw data. Given the importance of clarity of workflows and reproducibility of data in validating scientific results (1–3), we want to ensure that all of the relevant data contributing to an assembly are available for other researchers so that they can (i) reproduce the study's results, (ii) elaborate and incorporate the available data into other genome assemblies, or (iii) repurpose public data for use in alternative analyses. While many of these current best practices have been incorporated into the Instructions to Authors, in this opinion piece, we aim to provide a set of thematic ideas and examples behind certain instructions for authors to increase reproducibility across groups and utility for future users. We also highlight the fact that groups have proposed sets of standards for isolate genomes (4), 16S rRNA/18S rRNA/other amplicons (5), and single-cell amplified genomes (SAGs) and metagenome-assembled genomes (MAGs) (6) and that recommendations from those proposals are highly relevant and compatible with points raised in this editorial.

Citation Baltrus DA, Cuomo CA, Dennehy JJ, Dunning Hotopp JC, Maresca JA, Newton ILG, Rasko DA, Rokas A, Roux S, Stajich JE. 2019. Future-proofing your *Microbiology Resource Announcements* genome assembly for reproducibility and clarity. Microbiol Resour Announc 8:e00954-19. https://doi.org/10.1128/MRA.00954-19.

Address correspondence to David A. Baltrus, baltrus@email.arizona.edu.

The views expressed in this Editorial do not necessarily reflect the views of the journal or of ASM.

Published 5 September 2019

Strain provenance and culture conditions. Even before DNA extraction, it is important to document how particular isolates were isolated, cultured, and maintained. When and where was the strain isolated? What was the culture collection source? Has the strain been passaged since its isolation or acquisition from a culture collection? Was a single colony or plaque picked to amplify the culture? What kind of medium and growth conditions were used during growth of the organism prior to genome extraction? Deviations across these steps may not matter for the quality of proximate genome assemblies *per se*, but they can influence relevant measurements like estimation of the amount of polymorphisms compared to reference isolates and secondary patterns in which users might be interested, such as methylation status. There have been numerous studies demonstrating how common reference strains can accumulate changes simply because of independent maintenance across laboratories (e.g., see reference 7). Assembly of hypervariable genomic regions can also be significantly affected by polymorphisms that arise during culturing of strains prior to genomic extraction (8, 9). Other data, such as geographical coordinates, can be valuable to epidemiologists studying pathogen spread or evolutionary biologists studying isolation by distance. The more data provided relevant to the sample's provenance, the more useful the resource will be to future researchers.

Sample preparation. We often follow well-established protocols or use commercially available kits when extracting genomic DNA, and as such, it is commonplace and acceptable at MRA to provide references to specific methods or to state that procedures followed standard manufacturer protocols. However, these kits and protocols often include nonstandard or optional steps (e.g., addition of RNase); where possible, the inclusion of such steps should be documented in manuscripts because they can affect assembly quality. Likewise, it is valuable to include the type of kit used to create a sequencing library or prepare samples (Nextera/TruSeq, LSK108/RBK004, etc.), flow cell model or chemistry (FLO-MIN106, R9.4 pore, P6C4, etc.), if reads were multiplexed (and if so, what software was used to demultiplex or trim adaptors), and whether other DNA was sequenced in the same flow cell as part of the same run. Documentation of these steps can help reconcile biases that may influence genome assembly but also enable researchers to gauge the potential for contaminating reads to be incorporated in the reported genomes. When contracting with a commercial center or core, it is important to identify that center or core but also to verify that they will provide you with information required for publication. Such requirements currently include providing information about library construction methods, sequencing methods, sequencing platforms, and steps implemented in order to perform quality control for reads.

The sequencing of viruses may require additional information depending on the type of genome (linear or circular) and nucleic acid species (RNA or DNA). Different sample preparation strategies have different error profiles. For example, converting RNA genomes into cDNA prior to PCR amplification and Sanger sequencing has different strengths and weaknesses than those with applying sequence-independent, single-primer amplification (SISPA) and Illumina sequencing. Specifying the sample preparation strategies used can help other researchers understand the limitations of the sequencing effort.

Sequencing technologies. DNA sequencing technologies and assembly pipelines are rapidly changing. The best way to buffer against changes in genome assembly practices is to require that raw reads be deposited in a publicly available database, such as the NCBI Sequence Read Archive (SRA). Within reason, it is best if this information is posted in the least manipulated way so that researchers can derive the information in whatever way they would like. For instance, the removal of contamination of microbiome reads from a eukaryotic genome sequencing project could obscure secondary analysis of the microbiome of that eukaryote. It is especially critical that data underlying assemblies arising from sequencing reads generated by fast-changing technologies, like those generated through Oxford Nanopore devices, be extensively documented and accessible. To this point, since options for base calling from signals are rapidly

changing and improving for this platform, deposition of fast5 files into the SRA is critical for enabling future users to independently call bases or search for nucleotide modifications in the raw signals. As the software and algorithms for base calling are frequently changing, even if the assembly is based solely on the fastq reads that are produced by the MinKNOW pipeline, it is crucial to document versions of the base callers used within the pipeline (and all relevant parameters, since there are now options for "fast" or "high-accuracy" base calls). Last, given the variety of options currently available within the MinKNOW software, the selection of reads promoted to the assembly and the methods and cutoffs applied for filtering are critical to document (e.g., were they from the "pass" folder, or do they also include the "fail" folder?).

Towards fully reproducible genome assemblies. The more documentation that authors provide within each manuscript, the greater the possibility that results can be completely reproduced across labs and over time. We advocate for openness in terms of methods, sharing of all data, and deposition of relevant scripts described in manuscripts, and there are several ways that authors can achieve full transparency in these areas. We suggest that relevant and informative log files produced by software pipelines, which include information helpful for interpreting assembly metrics and pipeline dependencies, be made available through a publicly accessible data deposition archive like figshare or GitHub (https://guides.github.com/activities/citable-code/), linked to Zenodo (https://zenodo.org/), to enable documentation with digital object identifiers (DOIs). For instance, program packages like Unicycler (10) and Shovill (https://github.com/tseemann/shovill) output verbose log files that include parameters and versions of programs used in these packages, as well as inherent information such as how many rounds of Pilon (11) polishing each assembly underwent. Ultimately, the best solution possible is to post relevant information that can be used for benchmarking and quality control in accessible digital notebooks using programs like RMarkdown (12) or Jupyter (13) so that they are linked to DOIs that can be referenced in the manuscript.

REFERENCES

1. Madduri R, Chard K, D'Arcy M, Jung SC, Rodriguez A, Sulakhe D, Deutsch E, Funk C, Heavner B, Richards M, Shannon P, Glusman G, Price N, Kesselman C, Foster I. 2019. Reproducible big data science: a case study in continuous FAIRness. PLoS One 14:e0213013. https://doi.org/10.1371/journal.pone.0213013.
2. Yardımcı GG, Ozadam H, Sauria MEG, Ursu O, Yan K-K, Yang T, Chakraborty A, Kaul A, Lajoie BR, Song F, Zhan Y, Ay F, Gerstein M, Kundaje A, Li Q, Taylor J, Yue F, Dekker J, Noble WS. 2019. Measuring the reproducibility and quality of Hi-C data. Genome Biol 20:57. https://doi.org/10.1186/s13059-019-1658-7.
3. Poldrack RA, Gorgolewski KJ, Varoquaux G. 2019. Computational and informatic advances for reproducible data analysis in neuroimaging. Annu Rev Biomed Data Sci 2:119–138. https://doi.org/10.1146/annurev-biodatasci-072018-021237.
4. Field D, Garrity G, Gray T, Morrison N, Selengut J, Sterk P, Tatusova T, Thomson N, Allen MJ, Angiuoli SV, Ashburner M, Axelrod N, Baldauf S, Ballard S, Boore J, Cochrane G, Cole J, Dawyndt P, De Vos P, DePamphilis C, Edwards R, Faruque N, Feldman R, Gilbert J, Gilna P, Glöckner FO, Goldstein P, Guralnick R, Haft D, Hancock D, Hermjakob H, Hertz-Fowler C, Hugenholtz P, Joint I, Kagan L, Kane M, Kennedy J, Kowalchuk G, Kottmann R, Kolker E, Kravitz S, Kyrpides N, Leebens-Mack J, Lewis SE, Li K, Lister AL, Lord P, Maltsev N, Markowitz V, Martiny J, et al. 2008. The minimum information about a genome sequence (MIGS) specification. Nat Biotechnol 26:541–547. https://doi.org/10.1038/nbt1360.
5. Yilmaz P, Kottmann R, Field D, Knight R, Cole JR, Amaral-Zettler L, Gilbert JA, Karsch-Mizrachi I, Johnston A, Cochrane G, Vaughan R, Hunter C, Park J, Morrison N, Rocca-Serra P, Sterk P, Arumugam M, Bailey M, Baumgartner L, Birren BW, Blaser MJ, Bonazzi V, Booth T, Bork P, Bushman FD, Buttigieg PL, Chain PSG, Charlson E, Costello EK, Huot-Creasy H, Dawyndt P, DeSantis T, Fierer N, Fuhrman JA, Gallery RE, Gevers D, Gibbs RA, San Gil I, Gonzalez A, Gordon JI, Guralnick R, Hankeln W, Highlander S, Hugenholtz P, Jansson J, Kau AL, Kelley ST, Kennedy J, Knights D, Koren O, et al. 2011. Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. Nat Biotechnol 29:415–420. https://doi.org/10.1038/nbt.1823.
6. Bowers RM, Kyrpides NC, Stepanauskas R, Harmon-Smith M, Doud D, Reddy TBK, Schulz F, Jarett J, Rivers AR, Eloe-Fadrosh EA, Tringe SG, Ivanova NN, Copeland A, Clum A, Becraft ED, Malmstrom RR, Birren B, Podar M, Bork P, Weinstock GM, Garrity GM, Dodsworth JA, Yooseph S, Sutton G, Glöckner FO, Gilbert JA, Nelson WC, Hallam SJ, Jungbluth SP, Ettema TJG, Tighe S, Konstantinidis KT, Liu W-T, Baker BJ, Rattei T, Eisen JA, Hedlund B, McMahon KD, Fierer N, Knight R, Finn R, Cochrane G, Karsch-Mizrachi I, Tyson GW, Rinke C, Genome Standards Consortium, Lapidus A, Meyer F, Yilmaz P, Parks DH, et al. 2017. Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. Nat Biotechnol 35:725–731. https://doi.org/10.1038/nbt.3893.
7. Klockgether J, Munder A, Neugebauer J, Davenport CF, Stanke F, Larbig KD, Heeb S, Schöck U, Pohl TM, Wiehlmann L, Tümmler B. 2010. Genome diversity of Pseudomonas aeruginosa PAO1 laboratory strains. J Bacteriol 192:1113–1121. https://doi.org/10.1128/JB.01515-09.
8. Bao Z, Stodghill PV, Myers CR, Lam H, Wei H-L, Chakravarthy S, Kvitko BH, Collmer A, Cartinhour SW, Schweitzer P, Swingle B. 2014. Genomic plasticity enables phenotypic variation of Pseudomonas syringae pv. tomato DC3000. PLoS One 9:e86628. https://doi.org/10.1371/journal.pone.0086628.
9. Gaynor EC, Cawthraw S, Manning G, MacKichan JK, Falkow S, Newell DG. 2004. The genome-sequenced variant of Campylobacter jejuni NCTC 11168 and the original clonal clinical isolate differ markedly in colonization, gene expression, and virulence-associated phenotypes. J Bacteriol 186:503–517. https://doi.org/10.1128/jb.186.2.503-517.2004.
10. Wick RR, Judd LM, Gorrie CL, Holt KE. 2017. Unicycler: resolving bacterial genome assemblies from short and long sequencing reads. PLoS Comput Biol 13:e1005595. https://doi.org/10.1371/journal.pcbi.1005595.
11. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J, Young SK, Earl AM. 2014. Pilon: an integrated

tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One 9:e112963. https://doi.org/10.1371/journal.pone.0112963.
12. Allaire J, Cheng J, Xie Y, McPherson J, Chang W, Allen J, Wickham H, Atkins A, Hyndman R, Arslan R. 2016. RMarkdown: dynamic documents for R. R package version 1, 9010.
13. Kluyver T, Ragan-Kelley B, Pérez F, Granger BE, Bussonnier M, Frederic J, Kelley K, Hamrick JB, Grout J, Corlay S, Ivanov P, Avila D, Abdalla S, Willing C, Jupyter Development Team. 2016. Jupyter Notebooks—a publishing format for reproducible computational workflows, p 87–90. *In* Loizides F, Schmidt B (ed), Positioning and power in academic publishing: players, agents, and agendas. IOS Press, Clifton, VA.