


## LETTER TO THE EDITOR

# The Good, The Bad, and The Misleading: How to Improve the Quality of ‘Genome Announcements’?

Stefan Kusch,<sup>1</sup>  Niloofar Vaghefi,<sup>2</sup>  and Levente Kiss<sup>3,†</sup> 

<sup>1</sup> Unit of Plant Molecular Cell Biology, Institute for Biology I, RWTH Aachen University, Aachen, Germany

<sup>2</sup> Faculty of Science, University of Melbourne, Parkville, Victoria, Australia

<sup>3</sup> Centre for Crop Health, Institute for Life Sciences and the Environment, University of Southern Queensland, Toowoomba, Queensland, Australia

Accepted for publication 21 March 2023.

When comparing the requirements of diverse journals to publish microbial ‘Genome Reports,’ we noticed that some mostly focus on benchmarking universal single-copy orthologs scores as a quality measure, while the exclusion of possible contaminating sequences from genomic resources and the possible misidentification of the target microbes receive less attention. To deal with these quality issues, we suggest that DNA barcodes that are widely accepted for the identification of the target microbe species should be extracted from newly reported genome resources and included in phylogenetic analyses to confirm the identity of the sequenced microorganisms before Genome Reports are published. This approach, applied, for example, by the journal *IMA Fungus*, largely prevents the misidentification of the microbes that are targeted for whole-genome sequencing (WGS). In addition, contig similarity values, including GC content, remapping coverage of WGS reads, and BLASTN searches against the National Center for Biotechnology Information nucleotide database, would also reveal contamination issues. The values of these two recommendations to improve the publication criteria for microbial Genome Reports in diverse journals are demonstrated here through analyses of a draft genome published in *Molecular Plant-Microbe Interactions* and then retracted due to contaminations.

**Keywords:** benchmarking universal single-copy orthologs (BUSCO), genome assembly, obligate biotrophic plant pathogens, powdery mildews, quality control, whole-genome sequencing (WGS)

†Corresponding author: L. Kiss; [Levente.Kiss@usq.edu.au](mailto:Levente.Kiss@usq.edu.au)

**Funding:** S. Kusch was funded by the Deutsche Forschungsgemeinschaft (DFG) project number 274444799 (grant 861/14-2 awarded to R. Panstruga, RWTH Aachen University, Germany) in the context of the DFG-funded priority program SPP1819 “Rapid evolutionary adaptation – potential and constraints”. The paper is a result of a grant of the Australia-Germany Joint Research Co-Operation Scheme, and the program PPP Australia 2019 funded by the German Academic Exchange Service (DAAD) and also by the University of Southern Queensland.

**e-Xtra:** Supplementary material is available online.

The author(s) declare no conflict of interest.

Genome announcements are in vogue; currently, a number of prestigious research journals publish such papers, which report the completion of whole-genome sequencing (WGS) projects on certain organisms, the availability of the new data generated by the WGS project, and, sometimes, also a preliminary analysis of the results, without meeting all criteria of full-length research articles. Here, we focus on newly determined genomes of microbes that are regularly reported as ‘Genome Announcements,’ ‘Resource Announcements,’ ‘Genome Notes,’ or in other similar formats in a number of highly ranked journals. Table 1 lists examples of such journals in the broad field of microbiology. The table is a modified and updated version of an inventory provided by Smith (2017) in his analysis entitled “Goodbye genome paper, hello genome report: the increasing popularity of ‘genome announcements’ and their impact on science”—a great summary of this topic.

The requirements to publish microbial genome reports are quite different. Journals of The American Phytopathological Society (APS), for example, have focused on benchmarking universal single-copy orthologs (BUSCO) scores that refer to the completeness of genomic resources in terms of expected highly conserved genes. Currently, APS journals require that the sequenced strains are deposited in recognized herbaria or culture collections, that their identity was confirmed by phylogenetic analyses, pathogenicity tests, or other methods, singly or in combination, and that the genomic resources meet certain quality criteria (<https://apsjournals.apsnet.org/page/authorinformation>). Earlier, genomes of diverse plant pathogens and even genomes of different strains of the same species were published one after another as Resource Announcements in *Molecular Plant-Microbe Interactions* (MPMI), *Phytopathology*, and *Plant Disease* if BUSCO scores were satisfactory and the manuscripts met a few other general formal requirements. Since November 2022, these three APS journals do not accept submissions for Resource Announcements anymore; currently, *PhytoFrontiers* is the only APS journal that handles such submissions.

The journal *Genome Biology and Evolution* is more selective, and ‘Genome Reports’ are only published for “species not currently found in online databases, or where the previous sequence is of sub-standard quality” (a criterium listed on the journal webpage). *IMA Fungus*, a highly ranked journal of the publisher BioMed Central, produces another type of papers in their ‘Fungal Genomes’ section. Those papers appear twice a year, are multi-authored, and report several new, high-quality genomes for fungal species that do not have genomic resources available in public databases (Duong et al. 2021; Wingfield et al. 2022a



Copyright © 2023 The Author(s). This is an open access article distributed under the CC BY-NC-ND 4.0 International license.

and b). Most importantly, it is a requirement that Fungal Genomes papers in *IMA Fungus* include phylogenetic analyses of selected genes extracted from the newly reported genomes to confirm the identity of the sequenced strains.

Quality control of genome reports should be applied as widely as possible to avoid, for example, misclassification of sequences in reference databases and contamination of public genome assemblies with sequences from other organisms. These are two common issues in microbial genomics and have been the subject of many analyses (Breitwieser et al. 2019; Kryukov and Imanishi 2016; Lupo et al. 2021). The genomes of obligate biotrophic plant pathogens represent a special case because, in most cases, the sequenced DNA comes from non-axenic sources of the target microbial colonies. In those cases, in addition to the target plant pathogens (mostly fungi and oomycetes), the sequenced samples inevitably contain the DNA of the host plant tissues and also the DNA of other microbes that thrive inside or on the surface of the target colonies and their plant hosts (Panstruga and Kuhn 2019). Such contaminations in genome assemblies can lead to false conclusions; therefore, it is necessary to implement multiple methods and algorithms to identify and exclude contaminant sequences from the draft genomes of obligate biotrophic plant pathogens before making these public (Cornet et al. 2018; Kahlke and Ralph 2018; Kusch et al. 2020, 2022; Low et al. 2019; Wood et al. 2019; Zaccaron and Stergiopoulos 2021).

Recently, we analyzed all publicly available genomes of powdery mildew fungi (*Erysiphaceae*, Ascomycota) and produced the first comprehensive genome-scale phylogenetic analysis of this group of important obligate biotrophic plant pathogens (Vaghefi et al. 2022). Soon after our analysis was completed, the draft genome of the isolate NAFU1 of the grape powdery mildew fungus, *Erysiphe necator*, was published in *MPMI* (Zhang et al. 2021) and was subsequently retracted by the authors on 28 November 2022. As part of a follow-up study, we analyzed the retracted genome of *E. necator* isolate NAFU1 by performing sequence similarity searches (BLASTN) of all scaffolds against the GenBank nucleotide database and remapping the DNA sequencing data of the *E. necator* isolate C (GCA\_000798715.1), published by Jones et al. (2014), to the NAFU1 assembly. We used the internal transcribed spacer (ITS) sequence of the nuclear ribosomal DNA (nrDNA) of *Erysiphe necator* specimen MUMH 1835 (GenBank accession number LC175812) to extract potential nrDNA sequences in the genome

of NAFU1 and, further, conducted BLASTN searches of the extracted sequences against the National Center for Biotechnology Information (NCBI) nonredundant nucleotide database. This identified the presence of nrDNA sequences of *E. necator* within contig 52, while it also detected sequences within contigs 11, 18, 51, and 71 with identity or high similarity (>99.5%) to published nrDNA sequences of *Trichothecium roseum*, *Golubevia* spp., *Exobasidium* spp., and *Cladosporium* spp., respectively.

Additionally, to further demonstrate complications arising from the presence of contaminating sequences in NAFU1 genome assembly, we conducted phylogenetic analyses of representative ascomycetous genomes including *E. necator* isolate NAFU1 (Supplementary Table S1) based on 231 single-copy orthologous peptide sequences identified using Orthofinder v.2.5.1 (Emms and Kelly 2019) (concatenated alignment of 76,943 amino acids). This additional analysis made use of our previously established database of 751 single-copy orthologs identified in 24 powdery mildew genomes (Vaghefi et al. 2022). The phylogenetic analysis placed NAFU1 as a member of class *Sordariomycetes* and not *Leotiomycetes*, where powdery mildews belong (Fig. 1A). Similarly, BLASTN of all scaffolds revealed that only three, accounting for 232,275 bp of the assembly, were similar to powdery mildew sequences. Most of the sequences originated from fungi of the families *Hypocreaceae* and *Ustilaginaceae* (Fig. 1B), which belong to class *Sordariomycetes* and division *Basidiomycota*, respectively. Likewise, the DNA sequencing reads obtained from *E. necator* isolate C (Jones et al. 2014) exhibited low remapping percentage to the NAFU1 assembly, and most scaffolds had zero mapping coverage (Fig. 1B). Altogether, these data indicated that the retracted NAFU1 genome assembly mostly consisted of sequences from fungal contaminants and not *E. necator* or any other powdery mildew fungus.

Not surprisingly, our previous genome-scale phylogenetic analysis of family *Erysiphaceae* has also revealed that a number of published powdery mildew genome assemblies were contaminated with DNA sequences from non-target organisms; this was partly attributed to the difficulties of working with obligate biotrophic plant pathogens (Vaghefi et al. 2022). Highly contaminated or otherwise low-quality genome resources were excluded from our analyses (Vaghefi et al. 2022). Similar contaminations have been reported in the published genomes of other

**Table 1.** Examples of research journals that currently publish genome reports

Journal	Publisher	Article type	Notes
<i>PhytoFrontiers</i>	The American Phytopathological Society (APS)	Resource Announcements	Genome assemblies are accepted for any species of plant pathogens, including those with already published genomes, if the newly reported genomic resource is of much better quality. Papers are not expected to contain experimental data or address hypotheses, but a rationale is needed. Until November 1, 2022, <i>Molecular Plant-Microbe Interactions</i> , <i>Phytopathology</i> , and <i>Plant Disease</i> also accepted such submissions. Currently, <i>PhytoFrontiers</i> is the only APS journal that accepts this type of manuscript.
<i>Genome Biology and Evolution</i>	Oxford University Press	Genome Reports	Only genomes of organisms (not just microbes) that are of interest to the broad community of evolutionary biologists, and only species not currently found in online databases or where the previous sequence is of substandard quality are considered. A concise discussion of the insights to be gained from the new genome resource is required.
<i>IMA Fungus</i>	BioMed Central	IMA Genome	Multi-authored articles that report the first publicly available, high-quality genomes of fungal species that are of interest to the mycological community. A phylogenetic analysis of DNA species barcodes extracted from the newly reported genome assemblies must be included to support the identity of the sequenced fungal strains.
<i>Microbiology Resource Announcements</i>	American Society for Microbiology	Genome Sequences	Announcements of the availability of complete genome sequences or draft whole-genome sequences of any microbes to the scientific community.

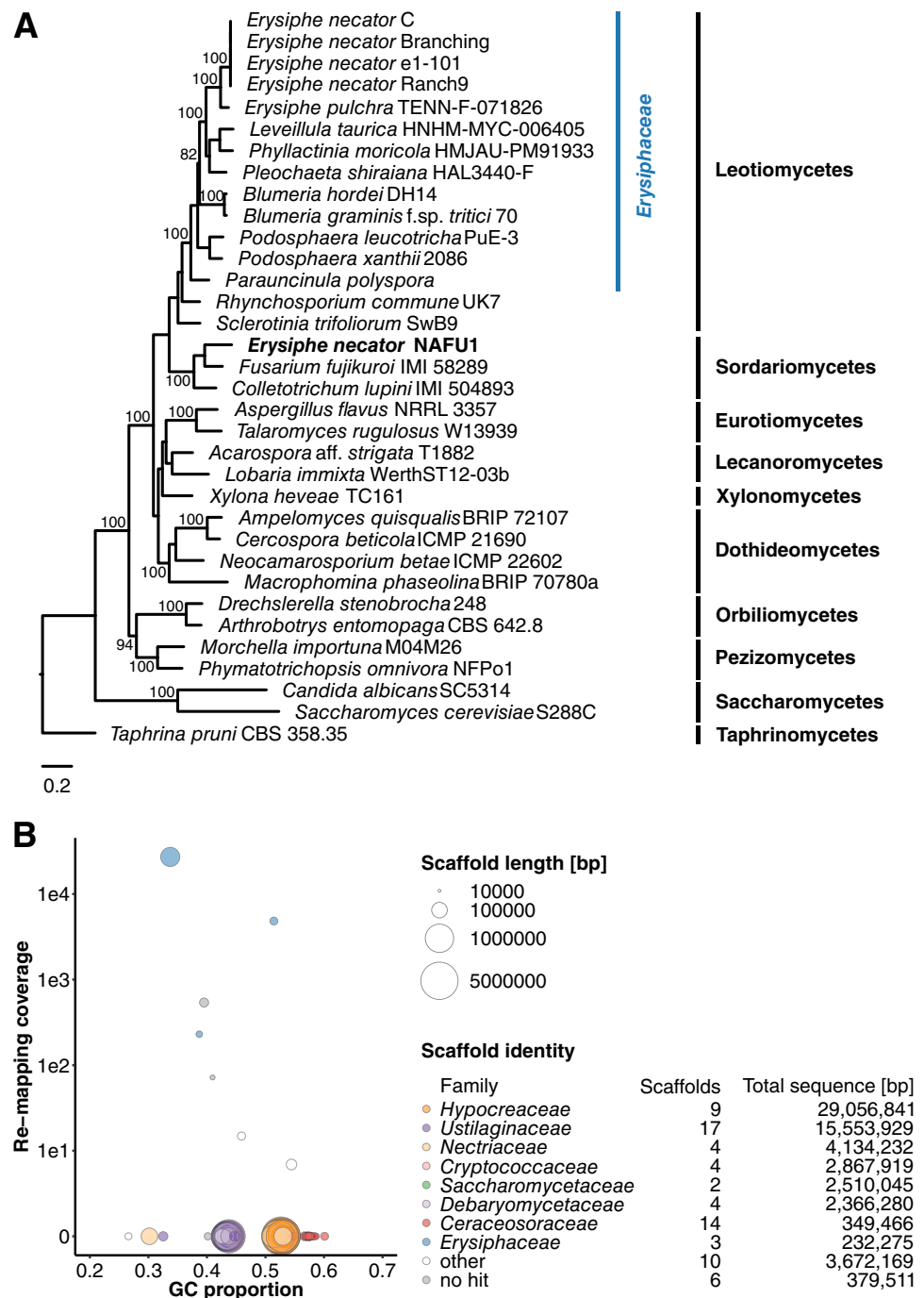
obligate biotrophs, e.g., *Albugo laibachii*, an oomycete infecting *Arabidopsis thaliana* (Zaccaron and Stergiopoulos 2021).

In our opinion, the examples provided above indicate that BUSCO values alone are not sufficient to warrant publication of genomic resources in research journals and public databases. We propose that genome contaminations should also be assessed through i) extraction and phylogenetic analyses of DNA barcode sequences from contigs and ii) contig similarity values, including GC content, remapping coverage of WGS reads, and BLASTN searches against the NCBI nucleotide database. Importantly, removal of contaminants should be explicitly listed as part of the basic criteria for quality assessment of genome assemblies.

Earlier, we recommended that DNA barcodes for the identification of species and genera, above all, nrDNA sequences in the

case of fungi, should also be used to assess the quality of genome assemblies for plant pathogens, in addition to the commonly used BUSCO values (Vaghefi et al. 2022). DNA barcodes, such as ITS sequences, that are widely accepted for the identification of the respective plant-pathogenic species and genera should be extracted from newly reported genome resources and included in a single- or multilocus phylogenetic analysis, depending on the taxonomic position of the sequenced specimen, to confirm the identity of the sequenced plant pathogens before accepting Genome Reports for publication. This approach, which is current practice at *IMA Fungus*, together with contig similarity scores detailed above should be used as quality measures of genome assemblies to make them more reliable and useful for the scientific community.

**Fig. 1.** Analyses of the genome assembly of *Erysiphe necator* isolate NAFU1, published by Zhang et al. (2021) in *Molecular Plant-Microbe Interactions* and then retracted by the authors on 28 November 2022. **A**, Maximum likelihood phylogenetic tree based on a concatenated alignment of 231 orthologous protein sequences derived from *E. necator* isolate NAFU1 and 33 representative ascomycetous genomes after removal of ambiguously aligned regions, using Gblocks v.0.91b (Castresana 2000; Talavera and Castresana 2007). The tree was constructed using RAxML-NG v.1.0.1 (Kozlov et al. 2019) with 1,000 bootstrap replicates, under the LG+I+G4+F amino acid substitution model identified by ModelTest-NG v.0.1.6 (Darriba et al. 2020). Taxon labels include species names followed by specimen or strain accession numbers. *Taphrina pruni* CBS 358.35 was used as the outgroup. The scale bar represents 0.2 nucleotide substitutions per site. Tree and alignment were submitted to TreeBASE (29834). **B**, Results of a BLASTN search against the National Center for Biotechnology Information GenBank nucleotide database accessed in April 2022. *Erysiphe necator* strain C genome sequencing data (Jones et al. 2014) were remapped to the *E. necator* NAFU1 assembly (GCA\_016906895.1; shown as remapping coverage on the y axis), and the GC content of each contig was calculated (x axis). The dot plot was generated via BlobTools (Laetsch and Blaxter 2017). Dot sizes indicate the length of the respective scaffolds or contigs, and color indicates the identity of the scaffold at the fungal family level, as shown in the legend.



## Literature Cited

- Breitwieser, F. P., Perteu, M., Zimin, A. V., and Salzberg, S. L. 2019. Human contamination in bacterial genomes has created thousands of spurious proteins. *Genome Res.* 29:954-960.
- Castresana, J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol. Biol. Evol.* 17:540-552.
- Cornet, L., Meunier, L., Vlierberghe, M. V., Léonard, R. R., Durieu, B., Lara, Y., Misztak, A., Sirjacobs, D., Javaux, E. J., Philippe, H., Wilmotte, A., and Baurain, D. 2018. Consensus assessment of the contamination level of publicly available cyanobacterial genomes. *PLoS One* 13:e0200323.
- Darriba, D., Posada, D., Kozlov, A. M., Stamatakis, A., Morel, B., and Flouri, T. 2020. ModelTest-NG: A new and scalable tool for the selection of DNA and protein evolutionary models. *Mol. Biol. Evol.* 37:291-294.
- Duong, T. A., Aylward, J., Ametrano, C. G., Poudel, B., Santana, Q. C., Wilken, P. M., Martin, A., Arun-Chinnappa, K. S., De Vos, L., DiStefano, I., Grewe, F., Huhndorf, S., Lumbsch, H. T., Rakoma, J. R., Poudel, B., Steenkamp, E. T., Sun, Y., van der Nest, M. A., Wingfield, M. J., Yilmaz, N., and Wingfield, B. D. 2021. IMA Genome-F15. *IMA Fungus* 12:30.
- Emms, D. M., and Kelly, S. 2019. OrthoFinder: Phylogenetic orthology inference for comparative genomics. *Genome Biol.* 20:238.
- Jones, L., Riaz, S., Morales-Cruz, A., Amrine, K. C., McGuire, B., Gubler, W. D., Walker, M. A., and Cantu, D. 2014. Adaptive genomic structural variation in the grape powdery mildew pathogen, *Erysiphe necator*. *BMC Genom.* 15:1081.
- Kahlke, T., and Ralph, P. J. 2018. BASTA – Taxonomic classification of sequences and sequence bins using last common ancestor estimations. *Methods Ecol. Evol.* 10:100-103.
- Kozlov, A. M., Darriba, D., Flouri, T., Morel, B., and Stamatakis, A. 2019. RAXML-NG: A fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics* 35:4453-4455.
- Kryukov, K., and Imanishi, T. 2016. Human contamination in public genome assemblies. *PLoS One* 11:e0162424.
- Kusch, S., Németh, M. Z., Vaghefi, N., Ibrahim, H. M. M., Panstruga, R., and Kiss, L. 2020. A short-read genome assembly resource for *Leveillula taurica* causing powdery mildew disease of sweet pepper (*Capsicum annuum*). *Mol. Plant-Microbe Interact.* 33:782-786.
- Kusch, S., Vaghefi, N., Takamatsu, S., Liu, S. Y., Németh, M. Z., Seress, D., Frantzeskakis, L., Chiu, P. E., Panstruga, R., and Kiss, L. 2022. First draft genome assemblies of *Pleochaeta shiraiana* and *Phyllactinia moricola*, two tree-parasitic powdery mildew fungi with hemiendophytic mycelia. *Phytopathology* 112:961-967.
- Laetsch, D. R., and Blaxter, M. L. 2017. BlobTools: Interrogation of genome assemblies. *F1000Research* 6:1287.
- Low, A. J., Koziol, A. G., Manninger, P. A., Blais, B., and Carrillo, C. D. 2019. ConFindr: Rapid detection of intraspecies and cross-species contamination in bacterial whole-genome sequence data. *PeerJ* 7:e6995.
- Lupo, V., Van Vlierberghe, M., Vanderschuren, H., Kerff, F., Baurain, D., and Cornet, L. 2021. Contamination in reference sequence databases: Time for divide-and-rule tactics. *Front. Microbiol.* 12:755101.
- Panstruga, R., and Kuhn, H. 2019. Mutual interplay between phytopathogenic powdery mildew fungi and other microorganisms. *Mol. Plant Pathol.* 20:463-470.
- Smith, D. R. 2017. Goodbye genome paper, hello genome report: The increasing popularity of 'genome announcements' and their impact on science. *Brief. Funct. Genomics* 16:156-162.
- Talavera, G., and Castresana, J. 2007. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Syst. Biol.* 56:564-577.
- Vaghefi, N., Kusch, S., Németh, M. Z., Seress, D., Braun, U., Takamatsu, S., Panstruga, R., and Kiss, L. 2022. Beyond nuclear ribosomal DNA sequences: Evolution, taxonomy, and closest known saprobic relatives of powdery mildew fungi (Erysiphaceae) inferred from their first comprehensive genome-scale phylogenetic analyses. *Front. Microbiol.* 13:903024.
- Wingfield, B. D., Berger, D. K., Coetzee, M. P. A., Duong, T. A., Martin, A., Pham, N. Q., van den Berg, N., Wilken, P. M., Arun-Chinnappa, K. S., Barnes, I., Buthelezi, S., Dahanayaka, B. A., Durán, A., Engelbrecht, J., Feurtey, A., Fourie, A., Fourie, G., Hartley, J., Kabwe, E. N. K., Maphosa, M., Mensah, D. L. N., Nsibo, D. L., Potgieter, L., Poudel, B., Stukenbrock, E. H., Thomas, C., Vaghefi, N., Welgemoed, T., and Wingfield, M. J. 2022b. IMA Genome-F17. *IMA Fungus* 13:19.
- Wingfield, B. D., De Vos, L., Wilson, A. M., Duong, T. A., Vaghefi, N., Botes, A., Kharwar, R. N., Chand, R., Poudel, B., Aliyu, H., and Barbetti, M. J., 2022a. IMA Genome-F16. *IMA Fungus* 13:3.
- Wood, D. E., Lu, J., and Langmead, B. 2019. Improved metagenomic analysis with Kraken 2. *Genome Biol.* 20:257.
- Zaccaron, A. Z., and Stergiopoulos, I. 2021. Characterization of the mitochondrial genomes of three powdery mildew pathogens reveals remarkable variation in size and nucleotide composition. *Microb. Genom.* 7:000720.
- Zhang, X., Mu, B., Cui, K., Liu, M., Ke, G., Han, Y., Wu, Y., Xiao, S., and Wen, Y. Q. 2021. Genome sequence resource for *Erysiphe necator* NAFU1, a grapevine powdery mildew isolate identified in Shaanxi Province of China. *Mol. Plant-Microbe Interact.* 34:1446-1449.