Misidentification of genome assemblies in public databases: The case of Naumovozyma dairenensis and proposal of a protocol to correct misidentifications

Stavrou, A.A.; Mixão, V.; Boekhout, T.; Gabaldón, T.

Published in:
Yeast

DOI:
10.1002/yea.3303

Received: 4 October 2017 | Revised: 30 November 2017 | Accepted: 21 December 2017

DOI: 10.1002/yea.3303

RESEARCH ARTICLE

WILEY Yeast

Misidentification of genome assemblies in public databases: The case of *Naumovozyma dairenensis* and proposal of a protocol to correct misidentifications

Aimilia A. Stavrou[1,2†] | Verónica Mixão[3,4†] | Teun Boekhout[1,2‡] | Toni Gabaldón[3,4,5‡]

[1] Westerdijk Fungal Biodiversity Institute, 3584 Utrecht, The Netherlands

[2] Institute for Biodiversity and ecosystem Dynamics, University of Amsterdam, 1012WX, Amsterdam, The Netherlands

[3] Centre for Genomic Regulation, The Barcelona Institute of Science and Technology, Dr. Aiguader 88, Barcelona 08003, Spain

[4] Universitat Pompeu Fabra, 08003 Barcelona, Spain

[5] Institució Catalana de Recerca i Estudis Avançats, Pg. Lluís Companys 23, 08010 Barcelona, Spain

Correspondence
Teun Boekhout, Westerdijk Fungal Biodiversity Institute, 3584 Utrecht, The Netherlands.
Email: t.boekhout@westerdijkinstitute.nl

Toni Gabaldón, Centre for Genomic Regulation, The Barcelona Institute of Science and Technology, Dr. Aiguader 88, Barcelona 08003, Spain.
Email: tgabaldon@crg.es

Funding information
H2020 Marie Skłodowska-Curie Actions, Grant/Award Number: 642095; Agència de Gestió d'Ajuts Universitaris i de Recerca, Grant/Award Number: SGR857; Agencia Nacional de Promoción Científica y Tecnológica, Grant/Award Number: BFU2015-67107; H2020 European Research Council, Grant/Award Number: ERC-2012-StG-310325

Abstract

Online sequence databases such as NCBI GenBank serve as a tremendously useful platform for researchers to share and reuse published data. However, submission systems lack control for errors such as organism misidentification, which once entered in the database can be propagated and mislead downstream analyses. Here we present an illustrating case of misidentification of *Candida albicans* from a clinical sample as *Naumovozyma dairenensis* based on whole-genome shotgun data. Analyses of phylogenetic markers, read mapping and single nucleotide polymorphisms served to correct the identification. We propose that the routine use of such analyses could help to detect misidentifications arising from unsupervised analyses and correct them before they enter the databases. Finally, we discuss broader implications of such misidentifications and the difficulty of correcting them once they are in the records.

KEYWORDS

Candida albicans, misidentification, *Naumovozyma dairenensis*, public databases

1 | INTRODUCTION

Part of our current work is the identification of species-specific genomic regions of human pathogenic yeasts. We have identified one such region, which is species-specific to *Candida albicans* and that shows a high similarity to its orthologue in *Candida dubliniensis*. The region of our interest is part of the ECE1 gene first described by Birse, Irwin, Fonzi, and Sypherdt (1993). It is easy to prove the species-specificity of this gene by aligning the firstly described ECE1 sequence in GenBank (accession number L17087). However, during our work, when performing a BLASTn search against the whole-genome shotgun (WGS) contigs database, as described in the 'Materials and methods'

†These authors contributed equally to this work. ‡These two authors share senior authorship.

section, one of the obtained hits was a region of the *Naumovozyma dairenensis* assembly associated with the BioProject PRJNA267549 (Roach, Burton, Lee, et al., 2015). This assembly was obtained by Roach et al. (2015), who applied WGS to identify pathogen isolates from patients in the clinical care unit. The study unveiled cryptic transmissions between patients and potential novel pathogenic strains. Among the strains identified, the one named 763_NDAI was claimed to correspond to *N. dairenensis*. As this species is phylogenetically distantly related to *C. albicans* and *C. dubliniensis* (Kurtzman, Fell, & Boekhout, 2011), the BLASTn hit seemed suspicious, prompting us to further investigate this finding. Here, we describe this investigation and discuss the implications of the presence of misidentified sequences in public databases.

2 | MATERIALS AND METHODS

2.1 | Detection of a possible misidentification

To verify the specificity of the ECE1 region of interest for our work to *C. albicans*, BLASTn was used (Altschul, Gish, Miller, Myers, & Lipman, 1990) with the following parameters: Database – Nucleotide collection, MegaBLAST; Max Target sequences – 500; and the rest of the parameters set to default. The same search was repeated, but changing the Database option to WGS (Organism Fungi: Taxid 4751). After this search, one of the obtained hits did not correspond to any *Candida* species, but to the *N. dairenensis* 763_NDAI strain assembly associated with the BioProject PRJNA267549 (Roach et al., 2015). Therefore, we decided to confirm whether this strain was correctly identified.

2.2 | Phylogenetic confirmation of the *N. dairenensis* assembly identification

To verify whether the 763_NDAI strain was correctly identified, we used phylogenetic markers traditionally used to identify filamentous fungi and yeasts, i.e. ribosomal DNA, elongation factor, DNA-directed RNA polymerase II and additional regions from the *N. dairenensis*-type strain CBS 421. These are the partial ribosomal DNA regions (small subunit ribosomal RNA gene, partial sequence; internal transcribed spacer 1, 5.8S ribosomal RNA gene, and internal transcribed spacer 2, complete sequence; and large subunit ribosomal RNA gene; accession number AJ229072), actin1 (accession number AF527937; Kurtzman & Robnett, 2003), DNA-directed RNA polymerase II (RPB2; accession number AF527908) and translation elongation factor 1 alpha (TEF1-*α*; accession number AF402046; Goddard & Burt, 1999). To determine the possible regions of 763_NDAI assembly corresponding to these regions, a BLASTn search of the sequences was performed against the WGS database selecting BioProject: PRJNA267549 (Roach et al., 2015). Then, we downloaded the top match of the aligned sequences and performed a BLASTn search against both nucleotide and WGS databases selecting organism Ascomycota Taxid: 4890. The matches of these sequences and the *N. dairenensis* strain 763_NDAI are shown in Table 1.

TABLE 1 BLASTn results for *Naumovozyma dairenensis* CBS 421 against *Naumovozyma dairenensis* strain 763_NDAI

Strain	Locus	Accession number	NCBI database	Query coverage	Identity
CBS 421	Partial rDNA	AJ229072	WGS	37%	88%
CBS 421	Actin1	AF527937	WGS	100%	86%
CBS 421	RPB2	AF527908	WGS	97%	67%
CBS 421	TEF1*α*	AF402046	WGS	99%	89%

WGS = Whole-genome shotgun.

2.3 | Genomic confirmation of the *N. dairenensis* assembly identification

To confirm that the WGS reads from 763_NDAI deposited in Sequence Read Archive (Leinonen, Sugawara, & Shumway, 2011; Roach et al., 2015) actually correspond to a *C. albicans* strain, as possibly indicated by the previous BLASTn hits, and not to *N. dairenensis*, we performed read mapping and single nucleotide polymorphism (SNP) calling on two available *C. albicans* reference genomes for strains SC5314 (van het Hoog, Rast, Martchenko, et al., 2007) and WO-1 (Butler, Rasmussen, Lin, et al., 2009) and to the reference genome of *N. dairenensis* CBS 421 (Gordon, Armisén, Proux-Wéra, et al., 2011). Briefly, 763_NDAI reads were inspected for their quality with FastQC, and reads with quality lower than 10 and/or size lower than 31 bp were filtered out. Then, they were mapped with BWA-MEM v0.7.15 (Li, 2013) against *C. albicans* SC5314 genome and *C. albicans* WO-1, both retrieved from Candida Genome Database (http://www.candidagenome.org/), the first one on 4 August 2017 and the second one on 12 January 2017, and *N. dairenensis* CBS 421 genome, retrieved from GenBank (https://www.ncbi.nlm.nih.gov/genbank/) on 16 January 2017. Picard v2.1.1 (http://broadinstitute.github.io/picard/) was used to sort reads by coordinates, mark the duplicates and get the mapping statistics. The mapping was inspected with IGV version 2.0.30 (Robinson, Thorvaldsdóttir, Winckler, et al., 2011). GATK v3.6 (McKenna, Hanna, Banks, et al., 2010) was used to determine variants. The HaplotypeCaller tool was set with --genotyping_mode DISCOVERY -stand_emit_conf 10 -stand_call_conf 30 -ploidy 2. A file containing only SNPs was generated with SelectVariants tool. The resulting file was filtered with VariantFiltration tool defining the following parameters: --clusterSize 5 --clusterWindowSize 20 --genotypeFilterName "heterozygous" --genotypeFilterExpression "isHet == 1" --filterName "bad_quality" -filter "QD < 2.0 || MQ < 40 || FS > 60.0 || HaplotypeScore > 13.0 || MQRankSum < -12.5 || ReadPosRankSum < -8.0". Mapping coverage was determined with SAMtools v0.1.18 (Li, Handsaker, Wysoker, et al., 2009). To calculate the number of SNPs per kilobase, only positions with one or more reads were considered for the genome size. Likewise, bedtools genomecov (Quinlan & Hall, 2010) was used to count the number of positions in the reference without any read mapped, and this number was subtracted from the total number of bases.

3 | RESULTS

3.1 | Phylogenetic confirmation of the *N. dairenensis* assembly identification

As mentioned before, after BLASTn of the ECE1 region used in our work, which is specific for *C. albicans*, all of the hits with significant scores are firstly *C. albicans* and with lower scores *C. dubliniensis* when the Nucleotide Database is selected in the search criteria. When the selection criteria are switched to WGS, the results that appear are 38 BLASTn hits, 35 of which are *C. albicans* strains, two *C. dubliniensis* strains and the *N. dairenensis* 763_NDAI strain.

To investigate a possible misidentification, we performed BLASTn searches against NCBI database using 763_NDAI sequences corresponding to several widely used phylogenetic markers as queries. The results from the BLASTn analysis of all of the sequences, phylogenetic markers, showed 99–100% identity for *C. albicans* strains in the nucleotide database and 100% identity with the *N. dairenensis* strain 763_NDAI and other *C. albicans* strains in the NCBI WGS database (Altschul et al., 1990), suggesting that 763_NDAI strain may indeed be *C. albicans* (Table 2).

3.2 | Genomic confirmation of the *N. dairenensis* assembly identification

To confirm the results above with sequences from the whole genome of the putative misidentified strains we mapped raw reads from the 763_NDAI assembly to the reference genomes of *C. albicans* SC5314 (van het Hoog et al., 2007) and WO-1 (Butler et al., 2009) strains, and to the reference genome of *N. dairenensis* CBS 421 reference (Gordon et al., 2011). Only 1.2% of the reads mapped to the *N. dairenensis* CBS 421 reference, while 98% and 96.7% of the reads were successfully mapped to *C. albicans* SC5314 and WO-1 genomes, respectively. SNP calling on the *C. albicans* strains identified a larger divergence with SC5314 (8.71 SNPs/kb) and WO-1 (8.75 SNPs/kb), with 4.01 and 4.06 homozygous SNPs/kb, respectively. In any case, this latter divergence value is within the range of previously reported differences among *C. albicans* strains of different clades that show an average of 3.7 SNPs/kb and 4.26 homozygous SNPs/kb in pair-wise comparisons (Hirakawa, Martinez, Sakthikumar, et al., 2015).

4 | DISCUSSION

The BLASTn results presented in this work, as well as the comparative genomics analysis performed, point to a misidentification of 763_NDAI strain, which seems to belong to the species *C. albicans*. It is worth mentioning that *N. dairenensis* has never been reported as a human pathogen causing disease nor even as a human commensal. A search in Pubmed Central showed 31 articles, none of which refer to *N. dairenensis* as a commensal or pathogen. The type strain CBS 421 has been isolated from dried fruit and other strains have been found on maize (Kurtzman et al., 2011). On the contrary, *C. albicans* is one of the main commensal yeasts on humans (Mayer, Wilson, & Hube, 2013) and a major cause of human yeast infection (Brown, Denning, Gow, et al., 2012).

TABLE 2 Blastn results for aligned regions of *Naumovozyma dairenensis* strain 763_NDAI against the Nucleotide database

Locus	NCBI Database	Species	Query Coverage	Identity
Partial rDNA	Nucleotide	*C. albicans*	100%	100%
Actin1	Nucleotide	*C. albicans*	100%	99%
RPB2	Nucleotide	*C. albicans*	100%	100%
TEF1a	Nucleotide	*C. albicans*	100%	99%

After confirming our suspected misidentification we set out to correct it from the record. We thus contacted the editors of the journal in which the paper containing this misidentification was published, presenting our case. Upon further inspection on the identities of other strains in this paper, it was found by the journal that there were several other cases of misidentifications and they worked with the authors of the original article to correct them. Later on a correction of the paper was published (Roach, Burton, Lee, et al., 2017). However, the misidentified strain reported in this work was not included in the corrected list.

Even when formally published, misidentifications can take a long while to lead to correction in public databases. A recent example is the misidentification of the sequences of NRRL 62431 strain as *Penicillium aurantiogriseum* (Yang, Zhao, Barrero, et al., 2014), which was shown to be *Penicillium expansum* more than two years ago (Ballester, Marcet-Houben, Levin, et al., 2015), yet it was only recently corrected in the GenBank database. Another example is the case of *Geotrichum candidum* strain 3C (Polev, Bobrov, Eneyskaya, & Kulminskaya, 2014) that was more than one year ago inferred to be a misidentified Pezizomycotina species (Shen, Zhou, Kominek, et al., 2016), and it has not yet been corrected in public databases. This genome information was used in subsequent studies (Borisova, Eneyskaya, Bobrov, et al., 2015; Hittinger, Rokas, Bai, et al., 2015), which highlights how misidentifications in databases are propagated. In a recent article, Nguyen and Boekhout (2017) raise the question of misidentifications in hybrid strains by reporting the case of incorrect nomenclature for *Saccharomyces uvarum*, which was for several years reduced to a variety of *Saccharomyces bayanus*. As the same authors suggest, and although in 2005 there was already an indication that these should be considered two different species (Nguyen & Gaillardin, 2005), in some databases this misidentification remains. Indeed, at the time of writing this article, in the NCBI Assembly database *S. uvarum* MCYC 623 and its spore clone 623-6C appear as belonging to different species, as the first one is correctly labelled as *S. uvarum* (ASM16699v1; Kellis, Patterson, Endrizzi, Birren, & Lander, 2003) and the second is considered *S. bayanus* (ASM16703v1; Cliften, Sudarsanam, Desikan, et al., 2003).

Whole genome sequencing data are currently flooding the databases and they undeniably provide a tremendous amount of valuable information. However, as in this case, the need for careful quality control of automated analyses and of curated databases is raised once more. Unfortunately, although misidentifications like the ones we presented here are rather easy to detect with comparative analyses, as shown, the errors are difficult to eliminate from the databases. Therefore, these errors can be propagated to several other studies, thus compromising their quality. To avoid such

situations, it is important to include identification checkpoints in databases before making the sequences publicly available. For instance, in the case of WGS, the inspection of specific regions known as good phylogenetic markers could help solving the problem. Here we would like to propose a general standard practice (Box 1). It is true that there is published literature discussing the advantages and limitations of using particular regions as suitable phylogenetic markers for certain fungal groups (Capella-Gutierrez, Kauff, & Gabaldón, 2014; Schoch, Seifert, Huhndorf, et al., 2012). However, there are also alternatives proposed, as for *Penicillium* spp. cytochrome c oxidase subunit 1 and β-tubulin can be alternatives to the rDNA region (Frisvad & Samson, 2004). Therefore, we believe that for the majority of cases this solution could avoid erroneous cases such as the ones described here.

Box 1 Fast and reliable identification of fungi and yeasts in WGS data.

To avoid misidentification of organisms there are certain steps that can be followed that, in the majority of cases, will ensure the correct identification of an unknown organism in the databases assuming that the species has been described and that some genome information regarding it is available in public databases. Since the correction of entries usually takes a considerable amount of time, incorporating the use of good phylogenetic markers in standard practice for sequence submission to public databases will greatly assist to decrease the cases of misidentifications. For this procedure to be more efficient it is also required that description of new species is accompanied by deposition of the sequences of the selected markers in the databases. As general guidelines we suggest some checkpoints before acceptance of sequences:

- The source of the strain is crucial. The habitat where a strain is isolated from can give a first indication of the identity. If not, at least the information can be used to cross-reference the result (see example with *Naumovozyma dairenensis*–*Candida albicans* case).
- If a strain is available, its deposition in a public culture collection is essential.
- The D1–D2 domains of the large subunit ribosomal DNA (LSU-rDNA) or Internal Transcribed spacers (ITS) are to be used in the majority of cases as they are popular regions for identification and databases contain a considerable number of such sequences from a wide variety of organisms. Identification at the genus level is possible most of the times with a simple BLASTn search in NCBI database. When D1–D2 does not provide an absolute species identification, additional markers should be used. The literature offers alternatives to D1-D2 and ITS depending on the fungus or yeast species (Stielow, Lévesque, Seifert, et al., 2015; Vu, Groenewald, Szoke, et al., 2016).
- The choice of the most appropriate marker usually depends on the genus of the organism. The literature offers alternatives to D1–D2 and ITS, which are usually housekeeping genes. These alternative markers include, but are not limited to, actin1, DNA-directed RNA polymerase II, translation elongation factor 1 α- or β-tubulin. Recently a set of four widespread marker genes able to phylogenetically resolve species relationships in dikaryotic fungi has been proposed (Capella-Gutierrez et al., 2014).
- We envision that an automated procedure to detect and compare such marker regions in a submitted sequence could be included in the submission process of public databases and used to detect potential misidentifications directly upon submission. A warning issued to the submitter or to database curators, coupled with a provisional halt of the submission, would allow correcting clear-cut errors before the sequences are publicly available.

ACKNOWLEDGEMENTS

This work has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement no. H2020-MSCA-ITN-2014-642095, 'OPATHY'. T.G. acknowledges support from the Spanish Ministry of Economy and Competitiveness grant BFU2015-67107 cofounded by European Regional Development Fund; from the European Union and ERC Seventh Framework Programme (FP7/2007-2013) under grant agreement ERC-2012-StG-310325; from the Catalan Research Agency (AGAUR) SGR857; and from the CERCA Programme/Generalitat de Catalunya. The authors declare that there is no conflict of interest.

ORCID

Aimilia A. Stavrou http://orcid.org/0000-0003-0877-1742
Verónica Mixão http://orcid.org/0000-0001-6669-0161
Teun Boekhout http://orcid.org/0000-0002-0476-3609
Toni Gabaldón http://orcid.org/0000-0003-0019-1735

REFERENCES

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

Ballester, A. R., Marcet-Houben, M., Levin, E., Sela, N., Selma-Lázaro, C., Carmona, L., … Gabaldón, T. (2015). Genome, transcriptome, and functional analyses of *Penicillium expansum* provide new insights into secondary metabolism and pathogenicity. *Molecular Plant-Microbe Interactions*, *28*(3), 232–248. https://doi.org/10.1094/MPMI-09-14-0261-FI

Birse, C. E., Irwin, M. Y., Fonzi, W. A., & Sypherdt, P. S. (1993). Cloning and characterization of ECE1, a gene expressed in association with cell elongation of the dimorphic pathogen *Candida albicans*. *Infection and Immunity*, *61*(9), 3648–3655.

Borisova, A. S., Eneyskaya, E. V., Bobrov, K. S., Jana, S., Logachev, A., Polev, D. E., … Stahlberg, J. (2015). Sequencing, biochemical characterization, crystal structure and molecular dynamics of cellobiohydrolase Cel7A from *Geotrichum candidum* 3C. *FEBS Journal*, *282*, 4515–4537. https://doi.org/10.1111/febs.13509

Brown, G. D., Denning, D. W., Gow, N. A., Levitz, S. M., Netea, M. G., & White, T. C. (2012). Hidden killers: human fungal infections. *Science Translational Medicine*, *4*(165), 165rv13. https://doi.org/10.1126/scitranslmed.3004404

Butler, G., Rasmussen, M. D., Lin, M. F., Santos, M. A., Sakthikumar, S., Munro, C. A., … Cuomo, C. A. (2009). Evolution of pathogenicity and sexual reproduction in eight *Candida* genomes. *Nature*, *459*(7247), 657–662. https://doi.org/10.1038/nature08064

Capella-Gutierrez, S., Kauff, F., & Gabaldón, T. (2014). A phylogenomics approach for selecting robust sets of phylogenetic markers. *Nucleic Acids Research*, *42*(7), e54. https://doi.org/10.1093/nar/gku071

Cliften, P., Sudarsanam, P., Desikan, A., Fulton, B., Majors, J., Waterson, R., … Johnston, M. (2003). Finding functional features in *Saccharomyces* genomes by phylogenetic footprinting. *Science*, *301*(5629), 71–76. https://doi.org/10.1126/science.1084337

Frisvad, J. C., & Samson, R. A. (2004). Polyphasic taxonomy of *Penicillium* subgenus *Penicillium* – A guide to identification of food and air-borne terverticillate penicillia and their mycotoxins. *Studies in Mycology*, *49*, 1–173.

Goddard, M. R., & Burt, A. (1999). Recurrent invasion and extinction of a selfish gene. *Proceedings of the National Academy of Sciences of the United States of America*, *96*(24), 13880–13885.

Gordon, J. L., Armisén, D., Proux-Wéra, E., ÓhÉigeartaigh, S. S., Byrne, K. P., & Wolfe, K. H. (2011). Evolutionary erosion of yeast sex chromosomes by mating-type switching accidents. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(50), 20024–20029. https://doi.org/10.1073/pnas.1112808108

Hirakawa, M. P., Martinez, D. A., Sakthikumar, S., Anderson, M. Z., Berlin, A., Gujja, S., … Cuomo, C. A. (2015). Genetic and phenotypic intra-species variation in *Candida albicans*. *Genome Research*, *25*(3), 413–425. https://doi.org/10.1101/gr.174623.114

Hittinger, C. T., Rokas, A., Bai, F., Boekhout, T., Gonçalves, P., Jeffries, T. W., … Kurtzman, C. P. (2015). For the 'Genomes and Evolution' special issue of *Current Opinion in Genetics and Development*. *Current Opinion in Genetics and Development*, *35*, 100–109. https://doi.org/10.1016/j.gde.2015.10.008

Kellis, M., Patterson, N., Endrizzi, M., Birren, B., & Lander, E. S. (2003). Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature*, *423*(6937), 241–254. https://doi.org/10.1038/nature01644

Kurtzman, C. P., Fell, J. W., & Boekhout, T. (Eds.) (2011). *The yeasts: A taxonomic study* (5th ed., Vol. 2). Amsterdam: Elsevier.

Kurtzman, C. P., & Robnett, C. J. (2003). Phylogenetic relationships among yeasts of the '*Saccharomyces* complex' determined from multigene sequence analyses. *FEMS Yeast Research*, *3*(4), 417–432.

Leinonen, R., Sugawara, H., & Shumway, M. (2011). The Sequence Read Archive. *Nucleic Acids Research*, *39*(Database issue, D19–D21. https://doi.org/10.1093/nar/gkq1019

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. ArXiv, 1303.3997.

Li, H., Handsaker, B., Wysoker, A., et al. (2009). The Sequence Alignment/map (SAM) format and SAMtools. *Bioinformatics*, *25*, 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Mayer, F. L., Wilson, D., & Hube, B. (2013). *Candida albicans* pathogenicity mechanisms. *Virulence*, *4*(2), 119–128. https://doi.org/10.4161/viru.22913

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., … DePristo, M. A. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, *20*(9), 1297–1303. https://doi.org/10.1101/gr.107524.110

Nguyen, H. C., & Gaillardin, C. (2005). Evolutionary relationships between the former species *Saccharomyces uvarum* and the hybrids *Saccharomyces bayanus* and *Saccharomyces pastorianus*; reinstatement of *Saccharomyces uvarum* (Beijerinck) as a distinct species. *FEMS Yeast Research*, *5*(4–5), 471–483. https://doi.org/10.1016/j.femsyr.2004.12.004

Nguyen, H. V., & Boekhout, T. (2017). Characterization of *Saccharomyces uvarum* (Beijerinck, 1898) and related hybrids: Assessment of molecular markers that predict the parent and hybrid genomes and a proposal to name yeast hybrids. *FEMS Yeast Research*, *17*(2). https://doi.org/10.1093/femsyr/fox014

Polev, E., Bobrov, K. S., Eneyskaya, E. V., & Kulminskaya, A. A. (2014). Draft genome sequence of *Geotrichum candidum* strain 3C. *Genome Announcements*, *2*(5), e00956-14. https://doi.org/10.1128/genomeA.00956-14

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*, *16*(6), 841–842. https://doi.org/10.1093/bioinformatics/btq033

Roach, D. J., Burton, J. N., Lee, C., Stackhouse, B., Butler-Wu, S. M., & Cookson, B. T. (2015). A year of infection in the intensive care unit: Prospective whole genome sequencing of bacterial clinical isolates reveals cryptic transmissions and novel microbiota. *PLoS Genetics*, *11*(7), e1005413. https://doi.org/10.1371/journal.pgen.1005413

Roach, D. J., Burton, J. N., Lee, C., Stackhouse, B., Butler-Wu, S. M., & Cookson, B. T. (2017). Correction: A year of infection in the intensive care unit: Prospective whole genome sequencing of bacterial clinical isolates reveals cryptic transmissions and novel microbiota. *PLoS Genetics*, *13*(4), e1006724. https://doi.org/10.1371/journal.pgen.1006724

Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., & Mesirov, J. P. (2011). Integrative genomics viewer. *Nature Biotechnology*, *29*(1), 24–26. https://doi.org/10.1038/nbt.1754

Schoch, C. L., Seifert, K. A., Huhndorf, S., Robert, V., Spouge, J. L., Levesque, C. A., … Fungal Barcoding Consortium. (2012). Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for fungi. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(16), 6241–6246. https://doi.org/10.1073/pnas.1117018109

Shen, X., Zhou, X., Kominek, J., Kurtzman, C. P., Hittinger, C. T., Rokas, A. (2016). Reconstructing the backbone of the Saccharomycotina yeast phylogeny using genome-scale data. *G3*, *6*(12):3927–3939. https://doi.org/10.1534/g3.116.034744

Stielow, J. B., Lévesque, C. A., Seifert, K. A., Meyer, W., Iriny, L., Smits, D., … Robert, V. (2015). One fungus, which genes? Development and assessment of universal primers for potential secondary fungal DNA barcodes. *Persoonia*, *35*, 242–263. https://doi.org/10.3767/003158515X689135

van het Hoog, M., Rast, T. J., Martchenko, M., Grindle, S., Dignard, D., Hogues, H., … Magee, P. T. (2007). Assembly of the *Candida albicans* genome into sixteen supercontigs aligned on the eight chromosomes. *Genome Biology*, *8*, R52. https://doi.org/10.1186/gb-2007-8-4-r52

Vu, D., Groenewald, M., Szoke, S., Cardinali, G., Eberhardt, U., Stielow, B., … Robert, V. (2016). DNA barcoding analysis of more than 9 000 yeast isolates contributes to quantitative thresholds for yeast species and genera delimitation. *Studies in Mycology*, *85*, 91–105. https://doi.org/10.1016/j.simyco.2016.11.007

Yang, Y., Zhao, H., Barrero, R. A., Zhang, B., Sun, G., Wilson, I. W., … Hoffman, A. (2014). Genome sequencing and analysis of the paclitaxel-producing endophytic fungus *Penicillium aurantiogriseum* NRRL 62431. *BMC Genomics*, *15*(1), 1–14. https://doi.org/10.1186/1471-2164-15-69

How to cite this article: Stavrou AA, Mixão V, Boekhout T, Gabaldón T. Misidentification of genome assemblies in public databases: The case of *Naumovozyma dairenensis* and proposal of a protocol to correct misidentifications. *Yeast*. 2018;35:425–429. https://doi.org/10.1002/yea.3303