Data in Brief 40 (2022) 107799

Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data Article

Genome sequence data of the antagonistic soil-borne yeast *Cyberlindnera sargentensis* (SHA 17.2)

Maria Paula Rueda-Mejia[a], Lukas Nägeli[a], Stefanie Lutz[b], Raúl A. Ortiz-Merino[c], Daniel Frei[b], Jürg E. Frey[b], Kenneth H. Wolfe[c], Christian H. Ahrens[b,d], Florian M. Freimoser[a,*]

[a] *Agroscope, Research Division Plant Protection, Müller-Thurgau-Strasse 29, Wädenswil 8820, Switzerland*
[b] *Agroscope, Competence Division Method Development and Analytics, Müller-Thurgau-Strasse 29, Wädenswil 8820, Switzerland*
[c] *Conway Institute, University College Dublin, Dublin 4, Ireland*
[d] *SIB, Swiss Institute of Bioinformatics, Müller-Thurgau-Strasse 29, Wädenswil 8820, Switzerland*

ARTICLE INFO

Article history:
Received 30 November 2021
Accepted 3 January 2022
Available online 5 January 2022

Keywords:
Antagonism
Biocontrol
Genome assembly and annotation
Mechanism
Plant protection
Yeast

ABSTRACT

Cyberlindnera sargentensis strain SHA 17.2, isolated from a Swiss soil sample, exhibited strong antagonistic activity against several plant pathogenic fungi *in vitro* and was highly competitive against other yeasts in soil. As a basis for identifying the mechanisms underlying its strong antagonistic activity, we have sequenced the genome of *C. sargentensis* (SHA 17.2) by long- and short read sequencing, *de novo* assembled them into seven contigs/chromosomes and a mitogenome (total genome size 11.4 Mbp), and annotated 5455 genes. This high-quality genome is the reference for transcriptome and proteome analyses aiming at elucidating the mode of action of *C. sargentensis* against fungal plant pathogens. It will thus serve as a resource for identifying potential biocontrol genes and performing comparative genomics analyses of yeast genomes.

* Corresponding author.
E-mail address: florian.freimoser@agroscope.admin.ch (F.M. Freimoser).
Social media: (M.P. Rueda-Mejia), (L. Nägeli), (S. Lutz), (D. Frei), (F.M. Freimoser)

https://doi.org/10.1016/j.dib.2022.107799
2352-3409/© 2022 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)

© 2022 The Authors. Published by Elsevier Inc.
This is an open access article under the CC BY-NC-ND
license (http://creativecommons.org/licenses/by-nc-nd/4.0/)

Specifications Table

Subject	Agricultural Microbiology
Specific subject area	Genome analysis of a yeast that strongly antagonises fungal plant pathogens.
Type of data	High-quality draft genome sequence data, genome annotation, table and figure
How data were acquired	Genomic DNA sequencing by Oxford Nanopore Technologies (ONT), Illumina MiSeq, and PacBio platforms, *de novo* assembly
Data format	Raw data: annotated draft genome assembly Secondary data: table of annotated genes, the encoding proteins, and functional prediction
Parameters for data collection	Genomic DNA was extracted from a pure culture of *C. sargentensis* (SHA 17.2) using a phenol/chloroform protocol.
Description of data collection	**Sequencing:** Oxford Nanopore Technologies (ONT), Illumina MiSeq, PacBio **Assembly:** filtering using length cut-offs, *de novo* assembly of PacBio reads, scaffolding with long ONT reads, reference-based assembly of the mitogenome. **Annotation:** Yeast Genome Annotation Pipeline (YGAP) and KEGG Orthologs assignment with KofaKOALA.
Data source location	*Cyberlindnera sargentensis* SHA 17.2 was isolated from a fallow farmland soil sample collected near Wädenswil (47.223140 °N, 8.676699 °E, 470 m.a.s.l.), Switzerland. The strain is available at the Culture Collection of Switzerland under CCOS1011.
Data accessibility	The assembled genome is deposited at NCBI's Genbank under the BioProject PRJNA763105 and the accession numbers CP083464-CP083471 (https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA763105). Additional data (PacBio and ONT long read data; Illumina miSeq short read data; genome annotation) is available at https://dataverse.harvard.edu/dataverse/Csar_genome.
Related research article	Hilber-Bodmer, M., Schmid, M., Ahrens, C.H., Freimoser, F.M., 2017. Competition assays and physiological experiments of soil and phyllosphere yeasts identify *Candida subhashii* as a novel antagonist of filamentous fungi. BMC Microbiol. 17, 4. 10.1186/s12866-016-0908-z

Value of the Data

- The genome of *C. sargentensis* (SHA 17.2; 7 contigs/chromosomes plus mitogenome) is the basis for identifying the biocontrol mode of action of this strongly antagonistic yeast.
- The annotated genome sequence released here can be used by biologists, microbiologists or mycologists who study fundamental aspects of microbial interactions or who are interested in developing new and improved biocontrol applications. Bioinformaticians and genome biologists may include the genome in comparative analyses and evolutionary studies.
- The high-quality genome of *C. sargentensis* (SHA 17.2) presented here is a reference for functional genomics studies and represents the basis for potential biocontrol genes and similarly active biocontrol strains through genome mining.

1. Data Description

Cyberlindnera sargentensis (SHA 17.2; CCoS1011) was isolated from an agricultural soil sample collected near Wädenswil (47.223140 °N, 8.676699 °E, 470 m.a.s.l.) in Switzerland. The strain was identified based on the ITS sequence as the species hypothesis SH1545207.08FU, which is currently labelled as *Cyberlindnera sargentensis* (Wick. & Kurtzman) Minter [1–3]. The isolate was

Table 1
Overview of the final, nearly complete *C. sargentensis* (SHA 17.2) *de novo* genome assembly.

	Chromosomes					Scaffolds		
Contigs	I	II	III	IV	V	VI	VII	Mitogenome
Length [bp]:								
	2,886,691	2,560,583	1,739,817	1,341,035	1,204,786	1,140,646	438,574	66,400
PacBio > 5 kb:								
Coverage	53x	54x	58x	63x	64x	69x	64x	1418x
Mapped				99.82 %				
ONT > 20 kb:								
Coverage	4x	5x	5x	7x	6x	8x	8x	298x
Mapped				100 %				
Illumina 2 × 300 bp:								
Coverage	65x	67x	74x	79x	82x	97x	105x	64x
Mapped				99.13 %				
No. of telomere patterns 5':								
	20	18	24	22	21	0	18	
No. of telomere patterns 3':								
	48	40	34	26	65	32	0	**Not annotated**
No. of genes								
	1403	1254	841	650	560	542	205	
No. of tRNAs								
	49	44	14	16	22	21	1	

Comments:
• I, II, IV, and V: Complete
• III: Complete apart from 10 kb of scaffolded Ns at 670 kb
• Mitogenome: Complete, circular
• VI: First 10 kb consist of collapsed rRNA operons. Two copies are present. The coverage is, however, ~20x higher. Thus, there should be ~40 copies, which can only be resolved using very long reads.
• VII: Scaffolds VI and VII might be on the same chromosome since the telomeres are missing at one end, and thus, are not complete. They also have a very similar coverage.

one of the most strongly antagonistic yeasts against a range of saprophytic and plant pathogenic filamentous fungi (e.g., *Botrytis, Fusarium*, and *Monilinia* strains) and was also highly competitive against other yeasts in soil [2,4]. *Cyberlindnera sargentensis* (SHA 17.2) has thus been selected as a promising yeast for potential biocontrol applications and for further characterising the mechanisms responsible for the strong biocontrol phenotype.

The initial *de novo* assembly of the *C. sargentensis* (SHA 17.2) genome consisted of 13 contigs, which, after ONT scaffolding, extensive polishing and manual curation, were reduced to a total of seven chromosomes and one mitogenome (Table 1). In order to correctly assemble the mitogenome, a reference-based approach was followed (see Methods), which resulted in the assembly of the 66 kb circular mitogenome. No additional plasmids could be identified. The total genome size was 11'378'532 bp. Variant calling detected only 55 and 12 variants in the Illumina and PacBio data, respectively, which suggested that *C. sargentensis* SHA 17.2 is a haploid strain. This was confirmed by the presence of only the MATa1 and MATa2 genes (CYSA0D04350 and CYSA0D04340, respectively) and the flanking genes SLA2 (CYSA0D04360) and VPS75 (CYSA0D04330), which often adjoin yeast MAT loci [5,6]. *C. sargentensis* is thus a heterothallic species and the strain SHA 17.2 a haploid of the mating type a. Overall, the small

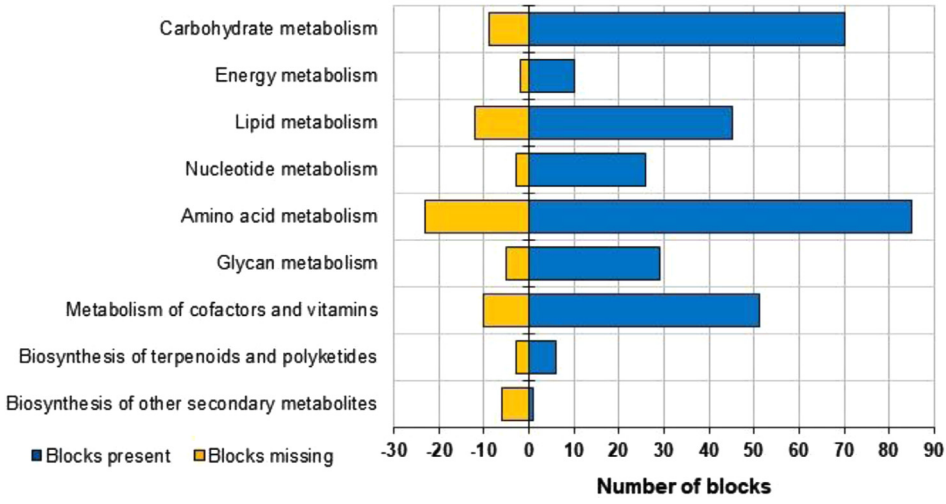

Fig. 1. Analysis of the *C. sargentensis* SHA 17.2 genome revealed many complete or nearly complete KEGG pathway modules. Out of the 5455 annotated protein coding genes, 3044 predicted *C. sargentensis* genes were matched with 3157 K numbers with a score above the predefined thresholds for individual KOs.

number of contigs and high coverage of the genome assembly (see Table 1) suggest the *C. sargentensis* SHA 17.2 genome to be of high quality and completeness.

The *C. sargentensis* nuclear genome contained 5455 protein coding genes and 167 tRNA genes. The mitochondrial genome was not annotated. Of all protein coding genes, 5019 sequences were annotated with at least one KEGG orthology identifier (KO identifier, K number). Overall, 3,157 K numbers with a score above the predefined thresholds for individual KOs were assigned to 3044 predicted *C. sargentensis* genes. Many KEGG pathway modules, functional units of gene sets in metabolic pathways, were complete or missed only few blocks as indicated by the KEGG Mapper Reconstruct tool [7] (Fig. 1). Based on the KofamKOALA KEGG Orthology analysis of the annotated genome, only one secondary metabolite biosynthesis gene (CYSA_0A07570; K06998, similar to a trans-2,3-dihydro-3-hydroxyanthranilate isomerase [EC:5.3.3.17]) was identified. However, the fungal antiSMASH v.6.0 online tool [8] identified two potential secondary metabolite clusters. The first represented a NRPS-like cluster predicted to consist of 15 genes that was localised on scaffold 1 (CYSA_0A11890-CYSA_0A12070). Furthermore, a predicted terpene cluster with seven genes was identified on scaffold 5 (CYSA_0E04890-CYSA_0E04950). Since antiSMASH uses different principles to predict genes, the annotations of the predicted secondary metabolite cluster genes were not identical to those from YGAP. Specific transcriptome and proteome analyses that are enabled by the *C. sargentensis* SHA 17.2 reference genome will help identifying a set of potential biocontrol genes by the strategy recently used for *Aureobasidium pullulans* [9].

2. Experimental Design, Materials and Methods

Genomic DNA was extracted using a phenol/chloroform extraction protocol. Oxford Nanopore Technologies (ONT) sequencing was carried out in-house. The ONT library was prepared using a 1D2 Sequencing Kit (SQK-LSK308) and sequenced on a FLO-MIN107 (R9.5) flow cell (all from Oxford Nanopore Technologies, Oxford, UK). One 2 × 300 bp Illumina paired end library was prepared in-house using the Nextera XT DNA kit and sequenced on a MiSeq platform (all from Illumina, Inc., San Diego, CA. USA). PacBio sequencing was carried out at the Functional Genomics Centre Zurich (FGCZ) on a Sequel machine (1 SMRT cell shared between three strains)

(PacBio, Menlo Park, CA. USA). Size selection was performed using the BluePippin system (Labgene Scientific, Châtel-St-Denis, Switzerland).

PacBio and ONT subreads were filtered with Filtlong (v.0.2.0) using a length cut-off of 5 kb and 20 kb, respectively. The Illumina reads were filtered and trimmed using trimmomatic (v0.39; parameters: phred 33, "LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36", only keep paired reads) [10]. The filtered PacBio reads were assembled using Flye (v.2.4; default parameters, except: estimated genome size of 11 Mb) [11], an assembly algorithm capable of resolving long, nearly identical repeat sequences [12]. Three short contigs were submitted to BLAT [13] and subsequently removed since they appeared spurious. The remaining 10 contigs were polished with the PacBio reads using 3 Arrow runs. The polished contigs were further scaffolded using the longer ONT reads (> 20 kb) and LRScaf (v.1.1.6). To correctly assemble the mitogenome, the mitogenome sequences of three *Cyberlindnera* strains (NC_022167.1, NC_022163.1, KC993181.1) were downloaded from NCBI and PacBio reads were individually mapped to the three references using minimap2 (set parameters: -a, -x map-pb). Mapping reads were filtered from the bam file using samtools (-F 4) and extracted into a fastq file using bam2fastq (v1.1.0). The reads were filtered by length (> 10 kb) and randomly subsampled (500 sequences) using awk to achieve a suitable coverage. The reads were assembled using Flye in plasmid mode (v.2.4; default parameters, except: estimated genome size of 50 kb, –plasmid) [11]. The circularity and completeness of the mitogenome were confirmed by mapping the PacBio reads to the start-aligned contig using minimap2 (set parameters: -a, -x map-pb) and visual inspection in the Integrative Genomics Viewer (IGV) [14]. All contigs were polished using the PacBio reads and 8 Arrow runs. The contigs were further polished using the Illumina reads and 3 Freebayes (v.1.2.0) [15] runs to correct potential small errors (e.g., homopolymer errors). The PacBio (> 5kb), ONT (> 20 kb) and Illumina reads were mapped to the polished contigs using minimap2 for PacBio (-x map-pb) and ONT (-x map-ont) and bwa for Illumina to verify the completeness and contiguity of the assembly by visual inspection in the IGV. PlasmidSpades [16] was run on the Illumina data in order to detect smaller plasmids. The mean telomere lengths (pattern "TGTGGTGTCTGGAT") could not be calculated using the Illumina reads and computel (v.1.2) [17]. The number of telomere patterns at both ends of each contig was thus counted manually (see Table 1). The ploidy level of the genome was estimated with the Illumina data by using PloidyNGS (v.3.1.2) [18] and nQuire [19]. Variants were called using the Illumina data and Freebayes (v.1.2.0; parameter: -C 5 (minimum count of observations supporting an alternate allele)) [15]; as well as the PacBio data and longshot (v.0.3.3) [20]. The variants were filtered using vcffilter and a quality cut-off of 20 (parameter: -f "QUAL > 20").

The *C. sargentensis* (SHA 17.2) genome was annotated using the Yeast Genome Annotation Pipeline (YGAP) [21]. Predictions were assessed for errors (i.e., internal stop codons, no ATG start codon) and manually corrected (indicated by the suffix "ed" in gene names). KEGG Orthologs (KOs; K numbers) were assigned to the predicted proteins by KofamKOALA [22]. The KEGG Mapper Reconstruct tool was used to assign the KOs to pathway modules [7].

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships, which have or could be perceived to have influenced the work reported in this article.

CRediT Author Statement

Maria Paula Rueda-Mejia: Investigation, Resources; **Lukas Nägeli:** Investigation, Resources; **Stefanie Lutz:** Software, Formal analysis; **Raúl A. Ortiz-Merino:** Software, Data curation, Formal analysis; **Daniel Frei:** Investigation, Resources; **Jürg E. Frey:** Resources, Supervision; **Kenneth H. Wolfe:** Software, Data curation, Supervision; **Christian H. Ahrens:** Conceptualization, Software, Supervision; **Florian M. Freimoser:** Conceptualization, Writing – review & editing, Supervision.

Acknowledgments

This work was funded by the Swiss National Science Foundation (SNSF) grant 31003A_175665/1 to FMF.

References

[1] K. Abarenkov, R.H. Nilsson, K.H. Larsson, I.J. Alexander, U. Eberhardt, S. Erland, K. Hoiland, R. Kjoller, E. Larsson, T. Pennanen, R. Sen, A.F.S. Taylor, L. Tedersoo, B.M. Ursing, T. Vralstad, K. Liimatainen, U. Peintner, U. Koljalg, The UNITE database for molecular identification of fungi - recent updates and future perspectives, New Phytol. 186 (2010) 281–285, doi:10.1111/j.1469-8137.2009.03160.x.

[2] M. Hilber-Bodmer, M. Schmid, C.H. Ahrens, F.M. Freimoser, Competition assays and physiological experiments of soil and phyllosphere yeasts identify *Candida subhashii* as a novel antagonist of filamentous fungi, BMC Microbiol. 17 (2017) 4 https://doi.org/10/ggd673.

[3] R.H. Nilsson, K.H. Larsson, A.F.S. Taylor, J. Bengtsson-Palme, T.S. Jeppesen, D. Schigel, P. Kennedy, K. Picard, F.O. Glockner, L. Tedersoo, I. Saar, U. Koljalg, K. Abarenkov, The UNITE database for molecular identification of fungi: handling dark taxa and parallel taxonomic classifications, Nucleic Acids Res. (2018) 2018/10/30, doi:10.1093/nar/gky1022.

[4] S. Gross, L. Kunz, D.C. Muller, A. Santos Kron, F.M. Freimoser, Characterization of antagonistic yeasts for biocontrol applications on apples or in soil by quantitative analyses of synthetic yeast communities, Yeast 35 (2018) 559–566 https://doi.org/10/gffv3k.

[5] S.J. Hanson, K.H. Wolfe, An evolutionary perspective on yeast mating-type switching, Genetics 206 (2017) 9–32 https://doi.org/10/ggkfrp.

[6] T. Krassowski, J. Kominek, X.-X. Shen, D.A. Opulente, X. Zhou, A. Rokas, C.T. Hittinger, K.H. Wolfe, Multiple Reinventions of Mating-type Switching during Budding Yeast Evolution, Curr. Biol. CB 29 (2019) 2555–2562 https://doi.org/e810/gmb35b.

[7] M. Kanehisa, Y. Sato, KEGG Mapper for inferring cellular functions from protein sequences, Protein Sci. Publ. Protein Soc. 29 (2020) 28–35 https://doi.org/10/gg3zj5.

[8] K. Blin, S. Shaw, A.M. Kloosterman, Z. Charlop-Powers, G.P. van Wezel, M.H. Medema, T. Weber, antiSMASH 6.0: improving cluster detection and comparison capabilities, Nucleic Acids Res. (2021) https://doi.org/10/gj28sg.

[9] M.P. Rueda-Mejia, L. Nägeli, S. Lutz, R.D. Hayes, A.R. Varadarajan, I.V. Grigoriev, C.H. Ahrens, F.M. Freimoser, Genome, transcriptome and secretome analyses of the antagonistic, yeast-like fungus *Aureobasidium pullulans* to identify potential biocontrol genes, Microb. Cell Graz Austria 8 (2021) 184–202 https://doi.org/10/gmkvdw.

[10] A.M. Bolger, M. Lohse, B. Usadel, Trimmomatic: a flexible trimmer for Illumina sequence data, Bioinformatics 30 (2014) 2114–2120 https://doi.org/10/f6cj5w.

[11] M. Kolmogorov, J. Yuan, Y. Lin, P.A. Pevzner, Assembly of long, error-prone reads using repeat graphs, Nat. Biotechnol. 37 (2019) 540–546 https://doi.org/10/gfzbrd.

[12] M. Schmid, D. Frei, A. Patrignani, R. Schlapbach, J.E. Frey, M.N.P. Remus-Emsermann, C.H. Ahrens, Pushing the limits of de novo genome assembly for complex prokaryotic genomes harboring very long, near identical repeats, Nucleic Acids Res. 46 (2018) 8953–8965, doi:10.1093/nar/gky726.

[13] W.J. Kent, BLAT–the BLAST-like alignment tool, Genome Res. 12 (2002) 656–664 https://doi.org/10/dpb3qz.

[14] H. Thorvaldsdóttir, J.T. Robinson, J.P. Mesirov, Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration, Brief. Bioinform. 14 (2013) 178–192 https://doi.org/10/f4sc43.

[15] E. Garrison, G. Marth, 2012. Haplotype-based variant detection from short-read sequencing. ArXiv12073907 Q-Bio.

[16] A. Bankevich, S. Nurk, D. Antipov, A.A. Gurevich, M. Dvorkin, A.S. Kulikov, V.M. Lesin, S.I. Nikolenko, S. Pham, A.D. Prjibelski, A.V. Pyshkin, A.V. Sirotkin, N. Vyahhi, G. Tesler, M.A. Alekseyev, P.A. Pevzner, SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing, J. Comput. Biol. 19 (2012) 455–477 https://doi.org/10/6zg.

[17] L. Nersisyan, A. Arakelyan, Computel: computation of mean telomere length from whole-genome next-generation sequencing data, PloS One 10 (2015) e0125201 https://doi.org/10/f7jtdt.

[18] R. Augusto Corrêa dos Santos, G.H. Goldman, D.M. Riaño-Pachón, ploidyNGS: visually exploring ploidy with Next Generation Sequencing data, Bioinformatics 33 (2017) 2575–2576 https://doi.org/10/gfz9cf.

[19] C.L. Weiß, M. Pais, L.M. Cano, S. Kamoun, H.A. Burbano, nQuire: a statistical framework for ploidy estimation using next generation sequencing, BMC Bioinf. 19 (2018) 122 https://doi.org/10/gfz9cg.

[20] P. Edge, V. Bansal, Longshot enables accurate variant calling in diploid genomes from single-molecule long read sequencing, Nat. Commun. 10 (2019) 4660 https://doi.org/10/ggnf6h.

[21] E. Proux-Wera, D. Armisen, K.P. Byrne, K.H. Wolfe, A pipeline for automated annotation of yeast genome sequences by a conserved-synteny approach, BMC Bioinf. 13 (2012) 237, doi:10.1186/1471-2105-13-237.

[22] T. Aramaki, R. Blanc-Mathieu, H. Endo, K. Ohkubo, M. Kanehisa, S. Goto, H. Ogata, KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold, Bioinformatics 36 (2020) 2251–2252 https://doi.org/10/ghk6qt.