Genomics Data 13 (2017) 41–43

Contents lists available at ScienceDirect

Genomics Data

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

Data in Brief

Genome sequence of the type strain CLIB 1764^T (= CBS 14374^T) of the yeast species *Kazachstania saulgeensis* isolated from French organic sourdough

Véronique Sarilar[a], Lieven Sterck[b,c], Saki Matsumoto[a], Noémie Jacques[a], Cécile Neuvéglise[d], Colin R. Tinsley[a], Delphine Sicard[e], Serge Casaregola[a,*]

[a] Micalis Institute, INRA, AgroParisTech, CIRM-Levures, Université Paris-Saclay, 78350 Jouy-en-Josas, France
[b] Ghent University, Department of Plant Biotechnology and Bioinformatics, Technologiepark 927, 9052 Ghent, Belgium
[c] VIB Center for Plant Systems Biology, 9052 Ghent, Belgium
[d] Micalis Institute, INRA, AgroParisTech, Université Paris-Saclay, 78350 Jouy-en-Josas, France
[e] Sciences pour l'œnologie, INRA, Supagro, Université de Montpellier, 34060 Montpellier, France

ARTICLE INFO

Keywords:
Saccharomycotina
Yeast
Sourdough
Kazachstania
Genome

ABSTRACT

Kazachstania saulgeensis is a recently described species isolated from French organic sourdough. Here, we report the high quality genome sequence of a monosporic segregant of the type strain of this species, CLIB 1764^T (= CBS 14374^T). The genome has a total length of 12.9 Mb and contains 5326 putative protein-coding genes, excluding pseudogenes and transposons. The nucleotide sequences were deposited into the European Nucleotide Archive under the genome assembly accession numbers FXLY01000001–FXLY01000017.

Specifications	
Organism/strain	*Kazachstania saulgeensis* strain CLIB 1764^T
Sex	N/A
Sequencer or array type	Illumina HiSeq 2500, mate pair libraries
Data format	Processed data: genome assembly and annotated embl files
Experimental factors	N/A
Experimental features	Genomic DNA extracted from pure yeast
Consent	N/A
Sample source location	Sourdough samples obtained from baker Michel Perrin at *Ferme des plants*, Saulgé, France (46° 20′ 27.08″ N 0° 53′ 12.21″ E)

1. Direct link to deposited data

https://www.ncbi.nlm.nih.gov/bioproject/PRJEB20516.

2. Introduction

The role of yeasts in bread making involves leavening the dough by fermenting carbon sources present in flour and producing aroma. In addition to the baker's yeast *Saccharomyces cerevisiae*, a number of other yeast species can be found in dough, in particular *Torulaspora delbrueckii*, *Wickerhamomyces anomalus* and *Pichia kudriavzevii* along with several members of the genus *Kazachstania*, such as *Candida humilis* (syn. *Candida milleri*, now *Kazachstania humilis*), *Kazachstania exigua*, less frequently *Kazachstania bulderi* and *Kazachstania unispora* [1,2]. A recent analysis of French organic sourdough revealed the presence a novel species, *Kazachstania saulgeensis* [2,3]. Here we report a high quality draft of the genome sequence of a monosporic segregant of the type strain of this species. The availability of the genome of *K. saulgeensis* will facilitate studies on the role of nonconventional yeasts in dough and the search for alternative baker's yeasts with interesting properties such as novel natural aromas.

3. Experimental design, materials and methods, results

Spore isolation from strain CLIB 1764^T grown on malt agar was performed as described in [4]. DNA from a single spore grown on YPD medium was prepared as previously described [4]. Preparation of two mate-pair libraries from the purified DNA and sequencing (Illumina HiSeq 2500 platform) was performed by BGI Genomics, Shenzhen,

* Corresponding author.
E-mail address: serge.casaregola@inra.fr (S. Casaregola).

http://dx.doi.org/10.1016/j.gdata.2017.07.003
Received 31 May 2017; Received in revised form 23 June 2017; Accepted 4 July 2017
Available online 04 July 2017

Table 1
Genome statistics for the strain CLIB 1764[T].

Attribute	CLIB 1764[T]
Genome size (bp)	12,935,755
Scaffolds > 10 kb	17
N50	1.37 Mb
G + C content	33%
Protein coding genes	5326
Pseudogenes	38
tRNA genes	197
LTR-retrotransposons (including pseudogenes)	15
Solo Long Terminal Repeats	278
DNA transposons (including. pseudogenes)	6

China. Two mate-pair libraries of 6-kbp insert size were sequenced, generating 6,055,467 read pairs of 100 bp and 5,496,657 read pairs of 125 bp. After trimming according to quality criteria with Trimmomatic [5], 21,095,636 reads were retained, leading to an apparent 190-fold coverage. The reads were assembled using Platanus, v1.2.1 [6] with default parameters. GapCloser v1.12 [7] was used to fill gaps where possible. The resulting assembly consisted of 3748 scaffolds with a maximum length of 2.96 Mb and with an N50 length of 1.37 Mb. The cumulative size was 13.99 Mb. The rDNA unit was assembled separately and manually integrated between the two scaffolds identified as being next to rDNA after mate-pair read mapping using BWA [8]. The resulting scaffold containing the rRNA locus was 0.89 Mb in size.

Annotation was performed on the 17 scaffolds larger than 10 kb (cumulative size of 12,935,755 bp, 32.5% GC content), whose size varied from 17.3 kb to 2.95 Mb (Table 1).

Based on the reference genomes of two related and well annotated, species belonging to the *Saccharomycetaceae*, *Saccharomyces cerevisiae* (http://www.yeastgenome.org/) and *Lachancea kluyveri* [9], a total of 5326 putative protein coding genes (CDS) and 38 pseudogenes were found using the Amadea Annotation transfer tool (Isoft, France). Functional annotation was performed based on protein similarity with *S. cerevisiae*. Coding sequences with no similarity to those in *S. cerevisiae* were annotated using the refseq and nr databases at NCBI. Further putative CDS were added after prediction of CDS longer than 150 aa with ORF Finder (http://www.ncbi.nlm.nih.gov/orffinder/) and blast analysis against the NCBI non redundant database, to yield a total of 5326 CDS (Table 1). Some of the gene models were manually curated on the ORCAE platform (http://bioinformatics.psb.ugent.be/orcae/; [10]) and visualized on GenomeView (http://genomeview.org; [11]). Interestingly, an arginase, whose gene had no equivalent in Saccharomycotina yeasts, but which presented strong sequence similarities with those of *Penicillium* is very likely the result of a horizontal gene transfer event.

One entire and one partial *Ty3*/gypsy retrotransposon were identified, together with 13 *Ty*-like pseudogenes. A total of 278 Long Terminal Repeats from retrotransposons were identified, belonging to at least 10 subfamilies. One of these subfamilies displays an unusual size of 714 bp, reminiscent of the long LTR found in *Kazachstania exigua* [12]. Members of two families of hAT DNA transposons, *Roamer* and *Rover* [13–15] with four and two elements respectively, were also identified; all were pseudogenes. A total of 197 tRNA were identified, using tRNAscan-SE v1.3.1 [16] (Table 1).

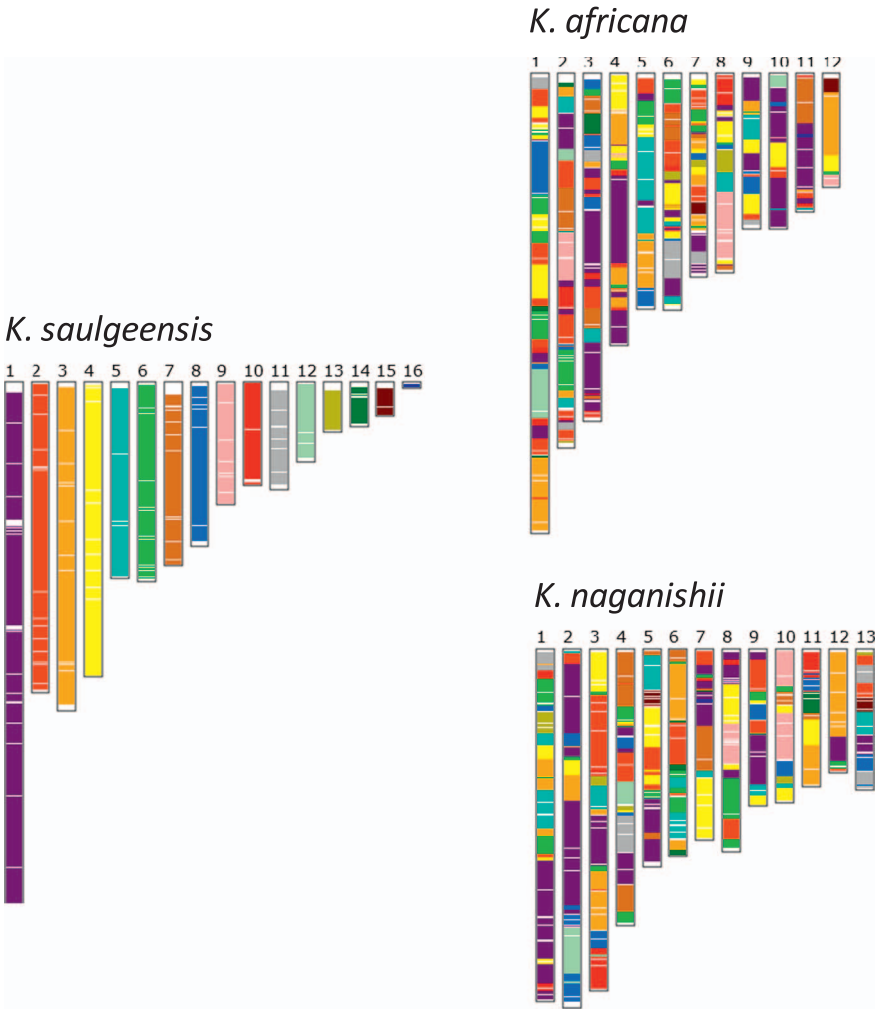

Fig. 1. Synteny blocks between the genomes of *K. saulgeensis* and two other *Kazachstania* species. Orthology relationships between genes from *K. africana*, *K. naganishii* and *K. saulgensis* were defined on the basis of bidirectional hits in a blastp comparison (reciprocal best hits) computed by SynChro [18]. The color attributed to the genes of a given *K. saulgeensis* scaffold is conserved for their counterparts in *K. africana* and *K. naganishii*.

We used the available genome of the type strain of two *Kazachstania* species, *Kazachstania africana* and *Kazachstania naganishii*, to investigate chromosome colinearity between *K. saulgeensis* and these species [17]. We examined the synteny based on the presence and order of orthologous genes using SynChro [18], with Delta = 4 to minimize artifactual synteny breaks. This showed that rearrangements that have occurred since the last common ancestor of *K. saulgeensis*, *K. africana* and *K. naganishii* are numerous and affect each scaffold equally (Fig. 1).

4. Nucleotide accession number

The genome sequences generated in this study are available from the European Nucleotide Archive under the genome assembly accession number GCA_900180425 and the scaffold accession range FXLY01000001–FXLY01000017. The genome can be browsed and searched at http://bioinformatics.psb.ugent.be/orcae/overview/Kasa.

Conflict of interest statement

The authors declare no conflict of interest.

Acknowledgments

This work was supported by an *Agence Nationale de la Recherche* grant ANR-13-ALID-0005 BAKERY (France). We are grateful to the INRA MIGALE bioinformatics platform (http://migale.jouy.inra.fr) for providing computational resources.

References

[1] L. De Vuyst, H. Harth, S. Van Kerrebroeck, F. Leroy, Yeast diversity of sourdoughs and associated metabolic properties and functionalities, Int. J. Food Microbiol. 239 (2016) 26–34.
[2] E. Lhomme, C. Urien, J. Legrand, X. Dousset, B. Onno, D. Sicard, Sourdough microbial community dynamics: an analysis during French organic bread-making processes, Food Microbiol. 53 (2016) 41–50.
[3] N. Jacques, V. Sarilar, C. Urien, M.R. Lopes, C.G. Morais, A.P. Uetanabaro, C.R. Tinsley, C.A. Rosa, D. Sicard, S. Casaregola, Three novel ascomycetous yeast species of the *Kazachstania* clade, *Kazachstania saulgeensis* sp. nov., *Kazachstania serrabonitensis* sp. nov. and *Kazachstania australis* sp. nov. Reassignment of *Candida humilis* to *Kazachstania humilis* f.a. comb. nov. and *Candida pseudohumilis* to *Kazachstania pseudohumilis* f.a. comb. nov, Int. J. Syst. Evol. Microbiol. 66 (2016) 5192–5200.
[4] S. Mallet, S. Weiss, N. Jacques, V. Leh-Louis, C. Sacerdot, S. Casaregola, Insights into the life cycle of yeasts from the CTG clade revealed by the analysis of the *Millerozyma* (*Pichia*) *farinosa* species complex, PLoS One 7 (2012) e35842.
[5] A.M. Bolger, M. Lohse, B. Usadel, Trimmomatic: a flexible trimmer for Illumina sequence data, Bioinformatics 30 (2014) 2114–2120.
[6] R. Kajitani, K. Toshimoto, H. Noguchi, A. Toyoda, Y. Ogura, M. Okuno, M. Yabana, M. Harada, E. Nagayasu, H. Maruyama, Y. Kohara, A. Fujiyama, T. Hayashi, T. Itoh, Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads, Genome Res. 24 (2014) 1384–1395.
[7] R. Luo, B. Liu, Y. Xie, Z. Li, W. Huang, J. Yuan, G. He, Y. Chen, Q. Pan, Y. Liu, J. Tang, G. Wu, H. Zhang, Y. Shi, C. Yu, B. Wang, Y. Lu, C. Han, D.W. Cheung, S.M. Yiu, S. Peng, Z. Xiaoqian, G. Liu, X. Liao, Y. Li, H. Yang, J. Wang, T.W. Lam, SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler, Gigascience 1 (2012) 18.
[8] H. Li, R. Durbin, Fast and accurate long-read alignment with Burrows-Wheeler transform, Bioinformatics 26 (2010) 589–595.
[9] N. Vakirlis, V. Sarilar, G. Drillon, A. Fleiss, N. Agier, J.P. Meyniel, L. Blanpain, A. Carbone, H. Devillers, K. Dubois, A. Gillet-Markowska, S. Graziani, N. Huu-Vang, M. Poirel, C. Reisser, J. Schott, J. Schacherer, I. Lafontaine, B. Llorente, C. Neuveglise, G. Fischer, Reconstruction of ancestral chromosome architecture and gene repertoire reveals principles of genome evolution in a model yeast genus, Genome Res. 26 (2016) 918–932.
[10] L. Sterck, K. Billiau, T. Abeel, P. Rouze, Y. Van de Peer, ORCAE: online resource for community annotation of eukaryotes, Nat. Methods 9 (2012) 1041.
[11] T. Abeel, T. Van Parys, Y. Saeys, J. Galagan, Y. Van de Peer, GenomeView: A next-generation genome browser, Nucleic Acids Res. 40 (2012) e12.
[12] E. Bon, C. Neuveglise, A. Lepingle, P. Wincker, F. Artiguenave, C. Gaillardin, S. Casaregola, Genomic exploration of the hemiascomycetous yeasts: 6. *Saccharomyces exiguus*, FEBS Lett. 487 (2000) 42–46.
[13] N. Rajaei, K.K. Chiruvella, F. Lin, S.U. Astrom, Domesticated transposase Kat1 and its fossil imprints induce sexual differentiation in yeast, Proc. Natl. Acad. Sci. U. S. A. 111 (2014) 15491–15496.
[14] V. Sarilar, C. Bleykasten-Grosshans, C. Neuveglise, Evolutionary dynamics of hAT DNA transposon families in Saccharomycetaceae, Genome Biol. Evol. 7 (2014) 172–190.
[15] J.L. Souciet, B. Dujon, C. Gaillardin, M. Johnston, P.V. Baret, P. Cliften, D.J. Sherman, J. Weissenbach, E. Westhof, P. Wincker, C. Jubin, J. Poulain, V. Barbe, B. Segurens, F. Artiguenave, V. Anthouard, B. Vacherie, M.E. Val, R.S. Fulton, P. Minx, R. Wilson, P. Durrens, G. Jean, C. Marck, T. Martin, M. Nikolski, T. Rolland, M.L. Seret, S. Casaregola, L. Despons, C. Fairhead, G. Fischer, I. Lafontaine, V. Leh, M. Lemaire, J. de Montigny, C. Neuveglise, A. Thierry, I. Blanc-Lenfle, C. Bleykasten, J. Diffels, E. Fritsch, L. Frangeul, A. Goeffon, N. Jauniaux, R. Kachouri-Lafond, C. Payen, S. Potier, L. Pribylova, C. Ozanne, G.F. Richard, C. Sacerdot, M.L. Straub, E. Talla, Comparative genomics of protoploid Saccharomycetaceae, Genome Res. 19 (2009) 1696–1709.
[16] T.M. Lowe, S.R. Eddy, tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence, Nucleic Acids Res. 25 (1997) 955–964.
[17] J.L. Gordon, D. Armisen, E. Proux-Wera, S.S. OhEigeartaigh, K.P. Byrne, K.H. Wolfe, Evolutionary erosion of yeast sex chromosomes by mating-type switching accidents, Proc. Natl. Acad. Sci. U. S. A. 108 (2011) 20024–20029.
[18] G. Drillon, A. Carbone, G. Fischer, SynChro: a fast and easy tool to reconstruct and visualize synteny blocks along eukaryotic chromosomes, PLoS One 9 (2014) e92621.