Data in brief 27 (2019) 104693

Contents lists available at ScienceDirect

Data in brief

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

Data Article

Draft genome sequence data of *Cercospora kikuchii*, a causal agent of Cercospora leaf blight and purple seed stain of soybeans

Francisco J. Sautua [a, 1], Sergio A. Gonzalez [b, 1], Vinson P. Doyle [c], Marcelo F. Berretta [d, e], Manuela Gordó [f], Mercedes M. Scandiani [g], Maximo L. Rivarola [b, d], Paula Fernandez [b, d, h], Marcelo A. Carmona [a, *]

[a] *Universidad de Buenos Aires, Facultad de Agronomía, Cátedra de Fitopatología, Buenos Aires, Argentina*
[b] *Instituto de Biotecnología, IABIMO, CICVyA, Instituto Nacional de Tecnología Agropecuaria, Hurlingham, Buenos Aires, Argentina*
[c] *Department of Plant Pathology and Crop Physiology, LSU AgCenter, Baton Rouge, LA, 70803, USA*
[d] *Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Buenos Aires, Argentina*
[e] *Instituto de Microbiología y Zoología Agrícola (IMyZA), Centro de Investigaciones en Ciencias Veterinarias y Agronómicas (CICVyA), Instituto Nacional de Tecnología Agropecuaria (INTA), Buenos Aires, Argentina*
[f] *Laboratorio Agrícola Río Paraná, San Pedro, Argentina*
[g] *Universidad Nacional de Rosario, Facultad de Ciencias Bioquímicas y Farmacéuticas, Centro de Referencia de Micología (CEREMIC), Rosario, Argentina*
[h] *Universidad Nacional de San Martín, Instituto de Investigaciones Biotecnológicas, Argentina*

ARTICLE INFO

Article history:
Received 1 September 2019
Received in revised form 24 September 2019
Accepted 15 October 2019
Available online 21 October 2019

Keywords:
Cercospora kikuchii
Draft genome
Next generation sequencing (NGS)
Cercospora leaf blight (CLB)
Purple seed stain (PSS)
Agriculture

ABSTRACT

Cercospora kikuchii (Tak. Matsumoto & Tomoy.) M.W. Gardner 1927 is an ascomycete fungal pathogen that causes Cercospora leaf blight and purple seed stain on soybean. Here, we report the first draft genome sequence and assembly of this pathogen. The *C. kikuchii* strain ARG_18_001 was isolated from soybean purple seed collected from San Pedro, Buenos Aires, Argentina, during the 2018 harvest. The genome was sequenced using a 2 × 150 bp paired-end method by Illumina NovaSeq 6000. The *C. kikuchii* protein-coding genes were predicted using FunGAP (Fungal Genome Annotation Pipeline). The draft genome assembly was 33.1 Mb in size with a GC-content of 53%. The gene prediction resulted in 14,856 gene models/14,721 protein coding genes. Genomic data of *C. kikuchii*

* Corresponding author.
E-mail address: carmonam@agro.uba.ar (M.A. Carmona).
[1] Authors having equal contribution.

https://doi.org/10.1016/j.dib.2019.104693

Bioinformatics
Fungal pathogens

presented here will be a useful resource for future studies of this pathosystem. The data can be accessed at GenBank under the accession number VTAY00000000 https://www.ncbi.nlm.nih.gov/nuccore/VTAY00000000.

Specifications table

Subject	Biology
Specific subject area	Bioinformatics (Genomics)
Type of data	Raw sequencing reads, draft genome assembly, gene prediction and phylogenetic position of *C. kikuchii* strain ARG_18_001
How data were acquired	Whole genome sequencing was performed using an Illumina NovaSeq 6000 sequencing system
Data format	Raw sequencing reads, draft genome assembly and gene prediction
Parameters for data collection	Reads were filtered and merged with Trimmomatic (v 0.39) and FLASH (v 1.2.11). The genome was assembled with Celera Assembler (v 8.3) and Spades (v 3.11.1). Gene prediction was performed with FunGAP (v 1.0.1), tRNAscan-SE (v 2.0.3), rnammer (v 1.2) and mfannot (v 1.35). Protein-coding gene annotation was performed with hmmsearch (v 3.1b2), ncbi-blast (v 2.2.25+) and Blast2GO (v 2.5) using the ragp R package (v 0.3.0.0001). RepeatMasker (v 4.0.9) was used to identify and filter repetitive regions.
Description of data collection	Strain ARG_18_001 was isolated from soybean seeds of variety DM62R63 sampled during the 2018 harvest that exhibited symptoms of purple seed stain.
Data source location	Samples were originally collected from Gobernador Castro, San Pedro, Buenos Aires, Argentina (33°39′26.37″S, 59°49′36.00″O)
Data accessibility	This Whole Genome Shotgun project has been deposited at DDBJ/ENA/GenBank under the accession VTAY00000000 https://www.ncbi.nlm.nih.gov/nuccore/VTAY00000000. The version described in this paper is version VTAY00000000.1 https://www.ncbi.nlm.nih.gov/nuccore/VTAY00000000.

Value of the Data

- The first draft genome of *Cercospora kikuchii* ARG_18_001.
- *C. kikuchii* is an important pathogen of soybean, but the biology of this fungus is poorly understood.
- Genomic data presented here will be a useful resource for the study of this pathosystem.
- This draft genome will help in the search for genetic resistance in soybean lines

1. Data

We present the draft genome assembly and gene prediction of the fungus *C. kikuchii*, causal agent of Cercospora leaf blight (CLB) and purple seed stain (PSS) of soybean. Recently, multi-locus phylogenetic studies confirmed that CLB and PSS is a disease complex caused by several *Cercospora* species. Phylogenetic analyses of cercosporoid fungi isolated from infected soybean in Argentina, Brazil and the USA determined that the species *C. kikuchii*, *C.* cf. *flagellaris* and *C.* cf. *sigesbeckiae* are causal agents of these diseases [1,2]. More recently, *C.* cf. *nicotianae* isolated from soybean leaves in Bolivia has been identified as a species in association with CLB [3]. A maximum-likelihood phylogenetic tree of *Cercospora* species was inferred in RAxML using seven nuclear loci, with data from isolate ARG_18_001 sliced from the genome assembly. The strain ARG_18_001 nested within the clade that includes other isolates of *C. kikuchii*, including the ex-type, with 97% bootstrap support (Fig. 1).

A total of 33,107,531 reads were assembled de novo, resulting in 136 scaffolds of at least 500 bp with the largest scaffold 3,211,885 bp and an N50 value of 898,622 bp. The mean coverage of the total assembly was 196.72-fold. The G + C content was 53.04%. The gene prediction resulted in

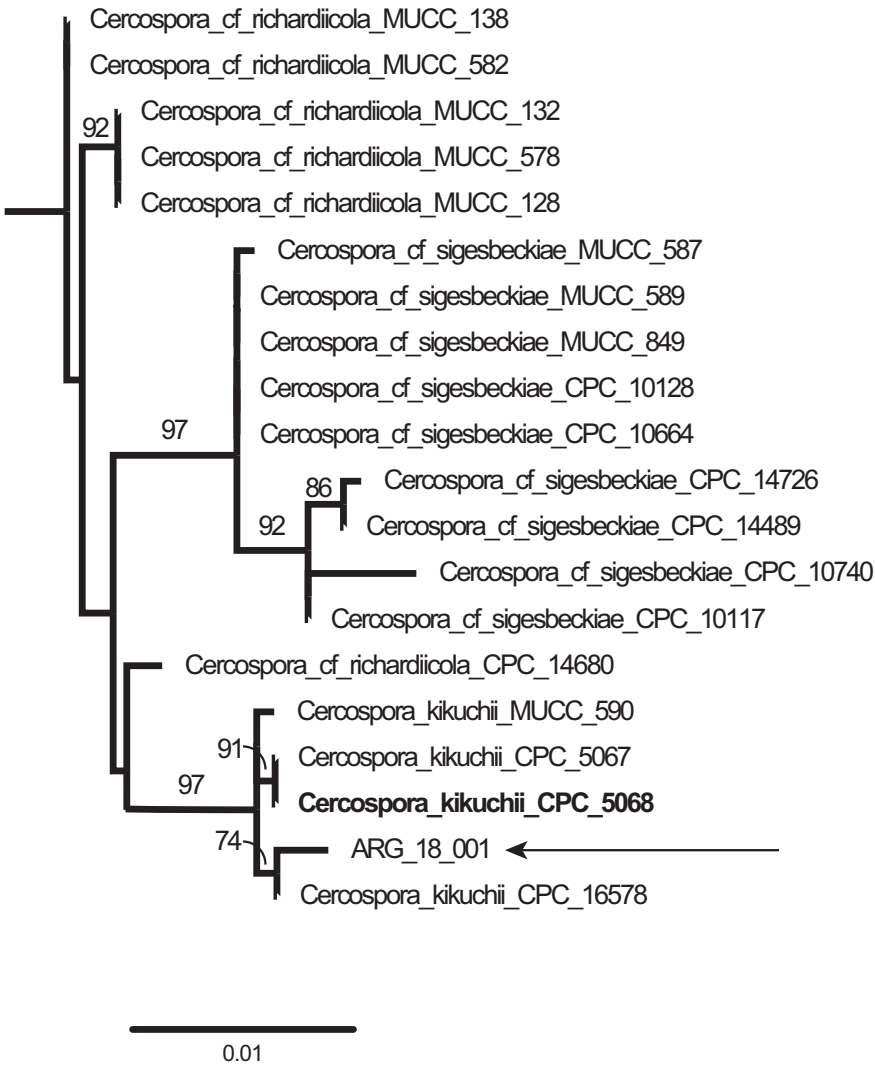

Fig. 1. Subtree from a maximum-likelihood phylogenetic analysis of *Cercospora* species. The complete phylogeny was inferred in RAxML assuming the GTRGAMMA model by integrating data sliced from the genome of ARG_18_001 with the following seven loci from 379 other isolates in Groenewald et al. (2013) and Bakhshi et al. (2018): *actA*, *cmdA*, *nrITS*, *gapdh*, *histone 3*, *tef1-alpha*, and *tub2*. The subtree that includes ARG_18_001 was pruned from the rest of the tree for ease of reference. Branches are labeled with bootstrap support values $\geq$ 70%. Bold font indicates the placement of the ex-type of *C. kikuchii*. Arrow indicates the placement of the isolate for which the genome was sequenced. The scale bar indicates the estimated number of substitutions per site.

14,856 gene models with 14,721 protein coding genes and 135 non coding RNAs, including the mitochondrial genome (Table 1). The distribution of protein annotations are summarized in Table 2, and Table 3 provides the summary statistics of the identified repetitive elements. The distribution of functional gene ontology (GO) terms from the annotated *C. kikuchii* ARG_18_001 genes are illustrated in Fig. 2. The distribution of species from the top BLAST hit of the predicted protein coding genes is shown in Fig. 3.

Table 1
Genome features of *C. kikuchii* strain ARG_18_001.

Features	*C. kikuchii* ARG_18_001
Assembled length	33,197,932
Scaffold length (≥ 50,000 bp)	32,541,287
Number of scaffolds (>500 bp)	136
Number of scaffolds (>1 kb)	107
Number of scaffolds (>50 kb)	71
Sequencing read coverage depth (fold)	196.72
GC-Content	53.04
No. of predicted protein-coding genes	14,721
Gene density (genes/Mb)	447.5
Average length of transcripts	1468.7
Average CDS length	1354.2
Average protein length	451.4
Average exon length	568.6
Average intron length	82.9
Spliced genes	9702 (66.0%)
Number of total introns	20,309
Median number of introns per gene	2.0
Number of total exons	35,010
Median number of exons per gene	2.0

Table 2
Genome annotation summary of *C. kikuchii* strain ARG_18_001.

Summary	Number
Number of protein-coding gene models	14,721
Number of models with BLAST hit	13,015 (88.4%)
Blast2GO annotation	6296 (42.8%)
PFAM annotation	5684 (38.6%)

Table 3
Summary of repetitive elements in the assembled genome of *C. kikuchii* strain ARG_18_001.

Summary	Number
Total of bases masked	178,815 (0.54%)
Number of simple repeats	3131
Number of low complexity repeats	358
Number of DNA transposons	68
Number of LTRs	2
Number of LINEs	254
Number of SINEs	21

2. Experimental design, materials, and methods

2.1. Genomic DNA extraction and sequencing

Cercospora kikuchii strain ARG_18_001 was isolated from a single conidium from soybean seeds of variety DM62R63 sampled that exhibited symptoms of purple seed stain during the 2018 harvest in San Pedro, Buenos Aires, Argentina. The isolation technique is described in [4]. This strain was deposited in the fungal culture collection of the Department of Plant Pathology, School of Agriculture, University of Buenos Aires (FAUBA, Argentina). Genomic DNA was isolated from hyphal tissue grown in potato dextrose broth for four days in darkness and constant agitation. The DNA extraction was carried out at the Institute of Microbiology and Agricultural Zoology (IMYZA -INTA) using a modified

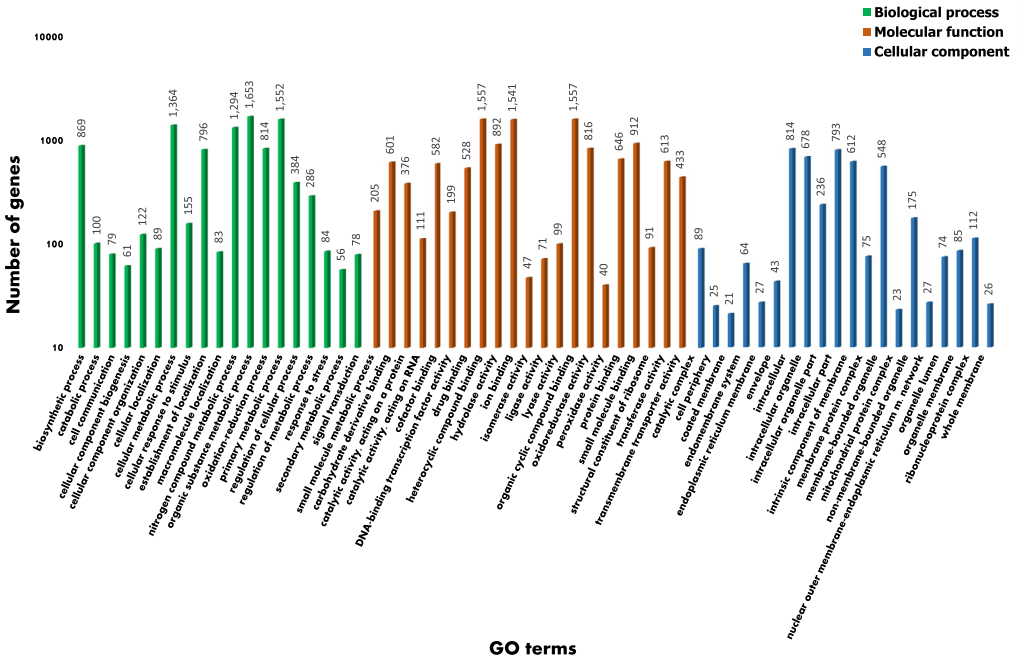

Fig. 2. Histogram representing the gene ontology distribution of the annotated *Cercospora kikuchii* ARG_18_001 genes. The functionally annotated genes were assigned to three main GO categories: Biological Process (BP), Molecular Function (MF) and Cellular Component (CC).

cetyltrimethylammonium bromide (CTAB) extraction protocol developed by [5]. Total DNA was quantified by fluorometry using a Picogreen dsDNA dye kit (Quant-iT, Invitrogen, by Life Technologies, CA, USA) with a Victor 3 plate reader.

Paired-end whole-genome shotgun libraries were constructed using the TruSeq Nano DNA (insert size 350 bp) library preparation kit following Illumina (San Diego, CA) protocols. Sequencing was performed using a NovaSeq 6000 sequencing system (Illumina) and yielded 65,202,278 reads.

2.2. Phylogenetic species identification

The isolate ARG_18_001 was identified by aligning seven nuclear loci (actin (*actA*), calmodulin (*cmdA*), nuclear ribosomal internal transcribed spacer region (*nrITS*), glyceraldehyde-3-phosphate dehydrogenase (*gapdh*), histone H3 (*his 3*), translation elongation factor 1-a (*tef1-alpha*) and beta tubulin (*tub2*)) with data from [6,7]. A maximum-likelihood phylogeny was then inferred in RAxML (Randomized Axelerated Maximum Likelihood) [8] assuming a GTRGAMMA model with *Septoria provencialis* CPC_12226 as an outgroup.

2.3. Genome assembly and annotation

Read trimming and filtering was performed using Trimmomatic [9] and merging of paired-end reads from shorter fragments was made using FLASH [10]. De novo assembly was carried out using the Celera Assembler [11] and then completed with Spades [12] using a wide range of k-mer values from 21 to 111 with a step of 2. The genome was annotated using FunGAP [13], tRNAscan-SE [14], rnammer [15] and MFannot (http://megasun.bch.umontreal.ca/cgi-bin/mfannot/mfannotInterface.pl) [16]. For predicting genes with FunGAP, the *C. kikuchii* ARG_18_001 genome assembly and the *C. beticola* 10.73.4 (Bioproject PRJNA294383) RNA-seq reads were used as inputs. To perform the

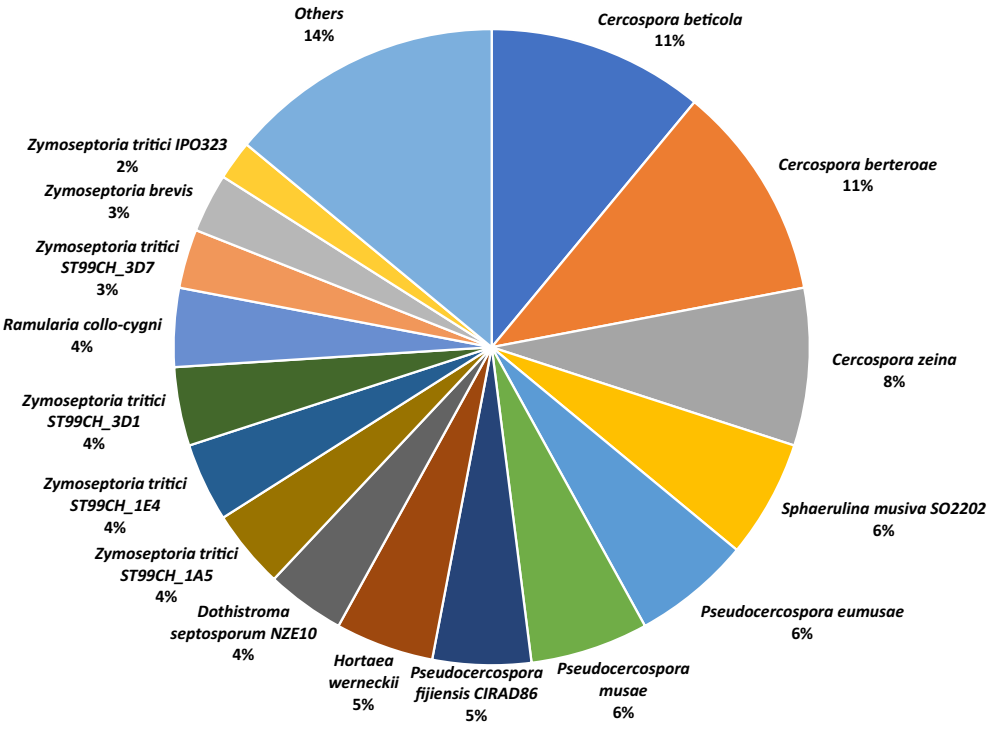

Fig. 3. Pie chart denoting the species distribution based on the top BLAST hit of the *Cercospora kikuchii* ARG_18_001 genes queried against the nr database with an E-value cut-off of 1E-10. The category "Others" includes species with less than 1% representation.

functional annotation, we used hmmsearch [17] against PFAM database (v32.0) (e-value cut off $\leq$ 10e-5) and BLASTP [18] (e-value cut off $\leq$ 10e-10) against the NCBI nr database. To assign Gene Ontology [19] terms we used Blast2GO [20] and pfam2go table (http://www.geneontology.org/external2go/pfam2go) with the ragp R package (https://rdrr.io/github/missuse/ragp/). The repetitive regions, including tandem repeats and transposable elements, were detected using the repeat identification tool RepeatMasker [21].

Acknowledgments

This work was financially supported by the University of Buenos Aires, Project UBACyT 20020170100147BA and partially by BASF Argentina S.A.

We specially thank Dr. Norma Paniego and the Bioinformatics Unit at IB/IABIMO INTA for technical assistance and support.

Conflict of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] A.P.G. Soares, E.A. Guillin, L.L. Borges, A.C.Td. Silva, Á.M.Rd. Almeida, P.E. Grijalba, A.M. Gottlie, B.H. Bluhm, L.O. Oliveira, More Cercospora species infect soybeans across the Americas than meets the eye, PLoS One 10 (2015), e0133495, https://doi.org/10.1371/journal.pone.0133495.

[2] S. Albu, R.W. Schneider, P.P. Price, V.P. Doyle, *Cercospora* cf. *flagellaris* and *Cercospora* cf. *sigesbeckiae* are associated with Cercospora leaf blight and purple seed stain on soybean in North America, Phytopathology 106 (2016) 1376–1385, https://doi.org/10.1094/PHYTO-12-15-0332-R.
[3] F.J. Sautua, J. Searight, V.P. Doyle, P.P. III Price, M.M. Scandiani, M.A. Carmona, The G143A mutation confers azoxystrobin resistance to soybean Cercospora leaf blight in Bolivia, Plant Health Prog. 20 (2019) 2–3, https://doi.org/10.1094/PHP-10-18-0060-BR.
[4] P. Price, M.A. Purvis, G. Cai, G.B. Padgett, C.L. Robertson, R.W. Schneider, S. Albu, Fungicide resistance in *Cercospora kikuchii*, a soybean pathogen, Plant Dis. 99 (2015) 1596–1603, https://doi.org/10.1094/PDIS-07-14-0782-RE.
[5] M.F. Berretta, R.E. Lecuona, R.O. Zandomeni, O. Grau, Genotyping isolates of the entomopathogenic fungus *Beauveria bassiana* by RAPD with fluorescent labels, J. Invertebr. Pathol. 71 (1998) 145–150, https://doi.org/10.1006/jipa.1997.4727.
[6] J.Z. Groenewald, C. Nakashima, J. Nishikawa, H.D. Shin, J.H. Park, A.N. Jama, M. Groenewald, U. Braun, P.W. Crous, Species concepts in Cercospora: spotting the weeds among the roses, Stud. Mycol. 75 (2013) 115–170, https://doi.org/10.3114/sim0012.
[7] M. Bakhshi, M. Arzanlou, A. Babai-Ahari, J.Z. Groenewald, P.W. Crous, Novel primers improve species delimitation in *Cercospora*, IMA Fungus 9 (2018) 299–332, https://doi.org/10.5598/imafungus.2018.09.02.06.
[8] A. Stamatakis, RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies, Bioinformatics 30 (2014) 1312–1313, https://doi.org/10.1093/bioinformatics/btu033.
[9] A.M. Bolger, M. Lohse, B. Usadel, Trimmomatic: a flexible trimmer for illumina sequence data, Bioinformatics 30 (2014) 2114–2120, https://doi.org/10.1093/bioinformatics/btu170.
[10] T. Magoc, S. Salzberg, FLASH: fast length adjustment of short reads to improve genome assemblies, Bioinformatics 27 (2011) 2957–2963.
[11] E.W. Myers, G.G. Sutton, A.L. Delcher, I.M. Dew, D.P. Fasulo, M.J. Flanigan, S.A. Kravitz, C.M. Mobarry, K.H. Reinert, K.A. Remington, E.L. Anson, R.A. Bolanos, H.H. Chou, C.M. Jordan, A.L. Halpern, S. Lonardi, E.M. Beasley, R.C. Brandon, L. Chen, P. J. Dunn, Z. Lai, Y. Liang, D.R. Nusskern, M. Zhan, Q. Zhang, X. Zheng, G.M. Rubin, M.D. Adams, J.C. Venter, A whole-genome assembly of *Drosophila*, Science 287 (2000) 2196–2204, https://doi.org/10.1126/science.287.5461.2196.
[12] A. Bankevich, S. Nurk, D. Antipov, A. Gurevich, M. Dvorkin, A.S. Kulikov, V. Lesin, S. Nikolenko, S. Pham, A. Prjibelski, A. Pyshkin, A. 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/10.1089/cmb.2012.0021.
[13] B. Min, I.V. Grigoriev, I.-G. Choi, FunGAP: fungal Genome Annotation Pipeline using evidence-based gene model evaluation, Bioinformatics 33 (18) (2017) 2936–2937, https://doi.org/10.1093/bioinformatics/btx353.
[14] P.P. Chan, T.M. Lowe, tRNAscan-SE: searching for tRNA genes in genomic sequences, Methods Mol. Biol. 1962 (2019) 1–14, https://doi.org/10.1007/978-1-4939-9173-0_1.
[15] K. Lagesen, P.F. Hallin, E. Rødland, H.H. Stærfeldt, T. Rognes, D.W. Ussery, RNAmmer: consistent and rapid annotation of ribosomal RNA genes, Nucleic Acids Res. 35 (2007) 3100–3108, https://doi.org/10.1093/nar/gkm160.
[16] N. Beck, B. Lang, MFannot, Organelle Genome Annotation Websever. http://megasun.bch.umontreal.ca/cgi-bin/mfannot/mfannotInterface.pl, 2010.
[17] S.R. Eddy, Accelerated profile HMM searches, PLoS Comput. Biol. 7 (10) (2011), e1002195, https://doi.org/10.1371/journal.pcbi.1002195.
[18] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, D.J. Lipman, Basic local alignment search tool, J. Mol. Biol. 215 (1990) 403–410, https://doi.org/10.1016/S0022-2836(05)80360-2.
[19] M. Ashburner, C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry, A.P. Davis, K. Dolinski, S.S. Dwight, J.T. Eppig, M.A. Harris, D.P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J.C. Matese, J.E. Richardson, M. Ringwald, G.M. Rubin, G. Sherlock, Gene ontology: tool for the unification of biology, Nat. Genet. 25 (2000) 25–29, https://doi.org/10.1038/75556.
[20] A. Conesa, S. Götz, J.M. García-Gómez, J. Terol, M. Talón, M. Robles, Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research, Bioinformatics 21 (Suppl. 18) (2005) 3674–3676, https://doi.org/10.1093/bioinformatics/bti610.
[21] M. Tarailo-Graovac, N. Chen, Using RepeatMasker to identify repetitive elements in genomic sequences, Curr. Protoc. Bioinform. 5 (2004) 4–10, https://doi.org/10.1002/0471250953.bi0410s25.