Genomics Data 4 (2015) 22–23

Contents lists available at ScienceDirect

Genomics Data

journal homepage: http://www.journals.elsevier.com/genomics-data/

Draft genome sequence of the intestinal parasite *Blastocystis* subtype 4-isolate WR1

Ivan Wawrzyniak [a], Damien Courtine [a], Marwan Osman [b], Christine Hubans-Pierlot [c], Amandine Cian [b], Céline Nourrisson [a], Magali Chabe [b], Philippe Poirier [a], Aldert Bart [d], Valérie Polonais [a], Pilar Delgado-Viscogliosi [b], Hicham El Alaoui [a], Abdel Belkorchia [a], Tom van Gool [d], Kevin S.W. Tan [e], Stéphanie Ferreira [c], Eric Viscogliosi [b,*], Frédéric Delbac [a,*]

[a] Clermont Université, Université Blaise Pascal-Université d'Auvergne-CNRS, UMR 6023 Laboratoire Microorganismes: Génome et Environnement, Clermont-Ferrand, France
[b] Institut Pasteur de Lille, Centre d'Infection et d'Immunité de Lille, Inserm U1019, CNRS UMR 8204, Université Lille, Nord de France, France
[c] Genoscreen Institut Pasteur, Lille, France
[d] Department of Medical Microbiology, Section Parasitology, Center for Infection and Immunity Amsterdam, Academic Medical Center, Amsterdam, The Netherlands
[e] Laboratory of Molecular and Cellular Parasitology, Department of Microbiology, Yong Loo Lin School of Medicine, National University of Singapore, Singapore

ARTICLE INFO

Article history:
Received 5 December 2014
Received in revised form 23 January 2015
Accepted 24 January 2015
Available online 3 February 2015

Keywords:
Blastocystis subtype 4-isolate WR1
Illumina-HiSeq
Whole genome
Annotation using Maker gene annotation pipeline

ABSTRACT

The intestinal protistan parasite *Blastocystis* is characterized by an extensive genetic variability with 17 subtypes (ST1–ST17) described to date. Only the whole genome of a human ST7 isolate was previously sequenced. Here we report the draft genome sequence of *Blastocystis* ST4-WR1 isolated from a laboratory rodent at Singapore.

Specifications	
Organism/cell line/tissue	*Blastocystis ST4*
Strain	WR1
Sequencer or array type	Illumina-HiSeq 2000
Data format	Processed
Experimental factors	Laboratory rodent and cultured axenically
Experimental features	Draft genome sequence of the intestinal parasite *Blastocystis* ST4-WR1 isolate
Consent	n/a
Sample source location	Clermont-Ferrand, France

Direct link to data

This Whole Genome Shotgun project has been deposited at DDBJ/EMBL/GenBank under the accession JPUL02000000 (http://www.ncbi.nlm.nih.gov/nuccore/JPUL00000000.2).

Experimental design, materials and methods

The stramenopile *Blastocystis* is a common anaerobic protist living in the digestive tract of several animal groups [1]. Its prevalence in human often exceeds 5% in industrialized countries [1] and can reach 100% in developing countries [2]. Although the role of *Blastocystis* as a human pathogen remains unclear, it has been associated with acute or chronic digestive disorders and some epidemiological surveys have suggested an association with irritable bowel syndrome (IBS) [3,4]. In patients with IBS, *Blastocystis* seems to be associated with a decrease of the fecal microbiota protective bacteria, *Bifidobacterium* sp. and *Faecalibacterium prausnitzii* [5]. The life cycle of the parasite is poorly documented. Among the parasitic forms described in the literature, the vacuolar stage which is maintained *in vitro* in axenic culture, is the most easily recognizable and the most frequently observed in stool samples. *Blastocystis* exhibits an extensive genetic diversity and seventeen subtypes (ST1–ST17) have been identified based on the gene coding for the small-subunit ribosomal RNA [6] among which the first nine are found in humans. The whole genome of a human *Blastocystis* ST7 isolate was previously sequenced. Briefly, it consists of an 18.8 Mbp nuclear

* Corresponding authors.
E-mail addresses: eric.viscogliosi@pasteur-lille.fr (E. Viscogliosi), frederic.delbac@univ-bpclermont.fr (F. Delbac).

http://dx.doi.org/10.1016/j.gdata.2015.01.009

Table 1
Genome statistics and intron features of *Blastocystis* ST4 and ST7.

	Blastocystis ST4	*Blastocystis* ST7
Genome assembly size	12.91 Mbp	18.8 Mbp
G + C content	39.6%	45.2%
Number of genes	5713	6021
Average gene size	1386 bp	1299 bp
Genes with introns	92.7%	84.6%
Average exon number per gene	5.06	4.58
Average length of introns (nt number)	33	50
Average length of proteins (aa number)	416	359
MLO genome size	27,815 bp	29,270 bp
MLO G + C content	21.94%	20.03%
Number of MLO genes	45	45

genome with 6020 predicted genes [7] and a circular genome of 29 kbp [8] located within mitochondria-like organelles (MLO). Other MLO genomes with conserved gene synteny have also been sequenced from *Blastocystis* ST1, ST3 and ST4 isolates [9,10]. Here we report the sequencing of the *Blastocystis* ST4-WR1 genome from an isolate of a laboratory rodent and cultured axenically [11]. Genomic DNA was isolated using a Qiagen DNeasy blood and tissue kit and sequencing was performed with the Illumina HiSeq 2000 system (Genoscreen, Lille, France). A total of 43.855.085 of 100-bp high quality paired-end reads were generated and were *de novo* assembled using the IDBA-ud algorithm [12]. The output was then scaffolded using SSPACE [13] and gaps were filled by Gapfiller software [14]. In total, 1301 scaffolds from 494 bp to 133,271 bp were obtained, with a scaffold N_{50} of 29,931 bp. The draft genome sequence of *Blastocystis* ST4 has a deduced total length of 12.91 Mbp and a G + C content of 39.7%. Assembly also provided a circular DNA molecule of 27,717 bp in size with a G + C content of 21.9% corresponding to the whole MLO genome sequence. Genes were carried out using the Maker gene annotation pipeline [15]. The Maker pipeline was set with the results of *ab initio* gene prediction algorithms Augustus [16] and SNAP [17], the 6020 protein-coding genes of *Blastocystis* ST7 [5], ESTs of both *Blastocystis* ST7 [5] and ST1 [18] and 414 manually-designed genes of the ST4-WR1 isolate. Basic information about the assembled genome and predicted genes are shown in Table 1. Gene functions were annotated by BLAST2GO [19] and BLAST analyses with NCBI (http://www.ncbi.nlm.nih.gov/). 183 tRNA were predicted using tRNAscan-SE 1.21 [20]. The preliminary annotation data revealed that *Blastocystis* ST4-WR1 nuclear genome harbors 5713 protein-coding genes. The presence of proteases was determined using BLAST against MEROPS database [21], and secreted proteases were identified using SIGNALP 3.0 [22] and WoLF PSORT [23]. Finally, OrthoMCL [24] was applied to compare both ST4 and ST7 genomes. This comparative analysis revealed that the ST4 genome contains less duplicated genes than ST7 and that more than 30% of ST4 genes have no ortholog in the ST7 genome at an *E* value cutoff of 10^{-5}. This also led to the identification of new candidate genes, in particular some potential virulence factors, including 20 secreted proteases that may be involved in the physiopathology of this parasite. Among these proteases, 7 seem to be specific to ST4 as no ortholog has been found in the ST7 genome. Sequencing and annotation of additional ST (ST1, ST2, ST3 and ST8) genomes are under progress and should be helpful for a better understanding of the genetic diversity, pathogenesis, metabolic potential and genome evolution of this highly prevalent human parasite.

Conflict of interest

Authors declare no conflict of interest.

Acknowledgments

This work was funded by grants from the French National Center for Scientific Research (CNRS), the INSERM, the Programme Orientations Stratégiques from the University of Lille 2 and the Institut Pasteur of Lille. MO was supported by a PhD fellowship from the Conseil National de la Recherche Scientifique and the Azm & Saade Association from Lebanon and AC by a PhD fellowship from the Pasteur Institute of Lille and the University of Lille 2.

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.gdata.2015.01.009.

References

[1] K.S. Tan, H. Mirza, J.D. Teo, B. Wu, P.A. Macary, Current views on the clinical relevance of *Blastocystis* spp. Curr. Infect. Dis. Rep. 12 (2010) 28–35.
[2] I. Wawrzyniak, P. Poirier, E. Viscogliosi, D. Meloni, C. Texier, F. Delbac, H.E. Alaoui, Blastocystis an unrecognized parasite: an overview of pathogenesis and diagnosis. Ther. Adv. Infect. Dis. 1 (2013) 167–178.
[3] P. Poirier, I. Wawrzyniak, C.P. Vivares, F. Delbac, H. El Alaoui, New insights into *Blastocystis* spp.: a potential link with irritable bowel syndrome. PLoS Pathog. 8 (2012) e1002545.
[4] D. El Safadi, L. Gaayeb, D. Meloni, A. Cian, P. Poirier, I. Wawrzyniak, F. Delbac, F. Dabboussi, L. Delhaes, M. Seck, M. Hamze, G. Riveau, E. Viscogliosi, Children of Senegal River Basin show the highest prevalence of *Blastocystis* sp. ever observed worldwide. BMC Infect. Dis. 14 (2014) 164.
[5] C. Nourrisson, J. Scanzi, B. Pereira, C. NkoudMongo, I. Wawrzyniak, A. Cian, E. Viscogliosi, V. Livrelli, F. Delbac, M. Dapoigny, P. Poirier, *Blastocystis* is associated with decrease of fecal microbiota protective bacteria: comparative analysis between patients with irritable bowel syndrome and control subjects. PLoS ONE 3 (2014) (2014) e111868.
[6] M.A. Alfellani, D. Taner-Mulla, A.S. Jacob, C. Atim Imeede, H. Yoshikawa, C.R. Stensvold, C.G. Clark, Genetic diversity of *Blastocystis* in livestock and zoo animals. Protist 164 (2013) 497–509.
[7] F. Denoeud, M. Roussel, B. Noel, I. Wawrzyniak, C. Da Silva, M. Diogon, E. Viscogliosi, C. Brochier-Armanet, A. Couloux, J. Poulain, B. Segurens, V. Anthouard, C. Texier, N. Blot, P. Poirier, G.C. Ng, K.S. Tan, F. Artiguenave, O. Jaillon, J.M. Aury, F. Delbac, P. Wincker, C.P. Vivares, H. El Alaoui, Genome sequence of the stramenopile *Blastocystis*, a human anaerobic parasite. Genome Biol. 12 (2011) R29.
[8] I. Wawrzyniak, M. Roussel, M. Diogon, A. Couloux, C. Texier, K.S. Tan, C.P. Vivares, F. Delbac, P. Wincker, H. El Alaoui, Complete circular DNA in the mitochondria-like organelles of *Blastocystis hominis*. Int. J. Parasitol. 38 (2008) 1377–1382.
[9] V. Perez-Brocal, C.G. Clark, Analysis of two genomes from the mitochondrion-like organelle of the intestinal parasite *Blastocystis*: complete sequences, gene content, and genome organization. Mol. Biol. Evol. 25 (2008) 2475–2482.
[10] C.R. Stensvold, M. Alfellani, C.G. Clark, Levels of genetic diversity vary dramatically between *Blastocystis* subtypes. Infect. Genet. Evol. 12 (2012) 263–273.
[11] X.Q. Chen, M. Singh, L.C. Ho, S.W. Tan, G.C. Ng, K.T. Moe, E.H. Yap, Description of a *Blastocystis* species from *Rattus norvegicus*. Parasitol. Res. 83 (1997) 313–318.
[12] Y. Peng, H.C. Leung, S.M. Yiu, F.Y. Chin, IDBA-UD: a *de novo* assembler for single-cell and metagenomic sequencing data with highly uneven depth. Bioinformatics 28 (2012) 1420–1428.
[13] M. Boetzer, C.V. Henkel, H.J. Jansen, D. Butler, W. Pirovano, Scaffolding pre-assembled contigs using SSPACE. Bioinformatics 27 (2011) 578–579.
[14] M. Boetzer, W. Pirovano, Toward almost closed genomes with GapFiller. Genome Biol. 13 (2012) R56.
[15] C. Holt, M. Yandell, MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. BMC Bioinf. 12 (2011) 491.
[16] K.J. Hoff, M. Stanke, WebAUGUSTUS — a web service for training AUGUSTUS and predicting genes in eukaryotes. Nucleic Acids Res. 41 (2013) W123–W128.
[17] I. Korf, Gene finding in novel genomes. BMC Bioinf. 5 (2004) 59.
[18] A. Stechmann, K. Hamblin, V. Pérez-Brocal, D. Gaston, G.S. Richmond, M. van der Giezen, C.G. Clark, A.J. Roger, Organelles that blur the distinction between mitochondria and hydrogenosomes. Curr. Biol. 18 (2008) 580–585.
[19] A. Conesa, S. Gotz, J.M. Garcia-Gomez, J. Terol, M. Talon, M. Robles, Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21 (2005) 3674–3676.
[20] 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.
[21] N.D. Rawlings, A.J. Barrett, A. Bateman, MEROPS: the database of proteolytic enzymes, their substrates and inhibitors. Nucleic Acids Res. 40 (2014) D343–D350.
[22] T.N. Petersen, S. Brunak, G. von Heijne, H. Nielsen, SignalP 4.0: discriminating signal peptides from transmembrane regions. Nat. Methods 8 (2014) 785–786.
[23] P. Horton, K.J. Park, T. Obayashi, N. Fujita, H. Harada, C.J. Adams-Collier, K. Nakai, WoLF PSORT: protein localization predictor. Nucleic Acids Res. 35 (2007) W585–W587.
[24] L. Li, C.J. Stoeckert Jr., D.S. Roos, OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. 13 (2003) 2178–2189.