Article

First Draft Genome of the Trypanosomatid

Herpetomonas muscarum ingenoplastis through

MinION Oxford Nanopore Technology and Illumina Sequencing

Claudia M. d’Avila-Levy1,2#, Bertrand Bearzatto3, Jérôme Ambroise3, Raphaël Helaers2, Anzhelika Butenko4,5, Vyacheslav Yurchenko5,6, Karina A. Morelli1,2, Helena L. C. Santos1, Pascal Brouillard2, Philippe Grellier8, Jean-Luc Gala3, Miikka Vikkula2

1 Coleção de Protozoários, Laboratório de Estudos Integrados em Protozoologia, Instituto Oswaldo Cruz, Fundação Oswaldo Cruz, 21040-360 Rio de Janeiro, Brazil; davila.levy@ioc.fiocruz.br
2 de Duve Institute, University of Louvain, B-1200, Brussels, Belgium; miikka.vikkula@uclouvain.be
3 Centre de TechnoLogies Moléculaires Appliquées, University of Louvain, B-1200, Brussels, Belgium; jean-luc.gala@uclouvain.be
4 Institute of Parasitology, Biology Centre, Czech Academy of Sciences, 37005 České Budějovice (Budweis), Czech Republic; rolando24@yandex.ru
5 Life Science Research Centre, Faculty of Science, University of Ostrava, 71000 Ostrava, Czech Republic; vyacheslav.yurchenko@osu.cz
6 Martsinovsky Institute of Medical Parasitology, Sechenov University, Moscow, 119435 Russia;
7 Instituto de Biología Roberto Alcântara Gomes, Departamento de Ecologia, Universidade Estadual do Rio de Janeiro, 20850-900, Rio de Janeiro, Brasil; kmorelli@ioc.fiocruz.br
8 Unité Molécules de Communication et Adaptation des Microorganisme (UMR 7245 CNRS MCAM), Muséum National d'Histoire Naturelle, Sorbonne Universités, 75005 Paris, France; grellier@mnhn.fr

* Correspondence: davila.levy@ioc.fiocruz.br (D.L.); Tel.: +55-21-2562-1014 (CMdL)

Abstract: We presented here the first draft genome sequence of the trypanosomatid Herpetomonas muscarum ingenoplastis. This parasite was isolated repeatedly in the black blowfly, Phormia regina. This is the first draft genome of a flagellate from the phylogenetically distinct clade of Trypanosomatidae.

Keywords: genome assembly; monoxenous trypanosomatids; insect trypanosomatids; Trypanosomatidae; whole genome

1. Introduction

The family Trypanosomatidae (Kinetoplastea: Trypanosomatida) comprises parasites of vertebrates, invertebrates or plants [1]. Chagas disease, leishmaniasis and human African trypanosomiasis are human diseases caused by Trypanosoma cruzi, Leishmania spp. and Trypanosoma brucei sensu lato, respectively [2]. These parasites affect about 22 million people worldwide and alternate their life cycle between an insect vector and a mammalian host [3]. Nonetheless, the largest biodiversity of this protist family is among trypanosomatids that usually carry out their entire life cycle inside insects [4-6]. Herpetomonas muscarum ingenoplastis was isolated and described by Rogers & Wallace in 1971 [7]. This parasite was capable of infecting flies from nine different genera, being the most prevalent Phormia. In artificial infections, it demonstrates high host specificity towards Phormia regina [7], which is a Palearctic fly found in North America and Northern Europe. Also known as ‘black blow fly’, it plays an important role in the ecosystem via carrion decomposition and nutrient recycling [8].
A BLAST analysis of the single available sequence of *H. muscarum ingenoplastis* (18S rRNA gene, GenBank Acc. number KX901631) revealed that it does not cluster with any other member of the genus *Herpetomonas*. Instead, its closest phylogenetic relatives (Trypanosomatidae spp. MCC-01, MCC-02, MCC-03, GMO-05, D44-1, G42, PNG60, and MCZ-14) form a separate group on the phylogenetic tree of trypanosomatids [9-11]. Here, we sequenced the whole genome of *H. muscarum ingenoplastis* combining MinION and Illumina.

2. Results

The Illumina sequencing yielded 100,372,731 reads, out of which 89.61% presented a Phred Q score of 30 or higher, and a mean quality score of 37.55. Regarding the MinION sequencing, the starting DNA presented a good quality with a DNA Integrity Number (DIN) of 9.1. After shearing, the majority of DNA (90% of the total) was composed of fragments from 3,208 bp to 46,456 bp, with an average size of 10,112 bp. Subsequently, a 1D sequencing library was run for approximately 43 h in a flow cell, generating a total of 2,402,163 reads. After base-calling, 88% of the total reads passed the mean quality score threshold of 7. The read N50 for those that passed the filter was 6,514, with 2637 reads longer than 20 kb, whereas the longest read was 54.8 kb.

The library generated using Canu consisted of 340 contigs, which were polished by the Illumina data using PILON. It resulted in a genome size of 35.09 Mb with an N50 of 375,483 bp, G+C content of 53.73%. The average coverages were 428X (MinION) and 270X (Illumina). The draft genome was aligned to *H. muscarum* reference genome by LastZ (v. 1.04.00) revealing that only 1.5% of the latter is covered by the assembly with an identity 80% or higher. This result underscores previous data from our research group, which indicate that this isolate is phylogenetically distant from all described trypanosomatids and must therefore be assigned to a new genus [11]. The automated annotation revealed a total of 8,619 genes. The long reads generated by the third-generation sequencing technologies, such as Oxford Nanopore Technologies (ONT), are particularly suitable to address the challenges associated with trypanosomatids genome, allowing direct determination of the full sequence of large clusters of repetitive sequences without collapsing them. In the fast changing field of long-read DNA sequencing, the Fiocruz Protist Collection decided to provide full genomic sequences of reference strains, as a strategic decision to boost science and promote Culture Collections [12].

3. Materials and Methods

*H. muscarum ingenoplastis* is cryopreserved at Fiocruz Protist Culture Collection (COLPROT) (http://colprot.fiocruz.br), voucher number COLPROT021. This specimen is also available at the American Type Culture Collection (ATCC30259). Flagellates were grown in a biphasic medium NNN/LIT (Novy-MacNeal-Nicolle/Liver Infusion Tryptose) supplemented with 10% fetal bovine serum. The genomic DNA was extracted using PureLink Genomic DNA mini kit (Invitrogen) from cells in the late logarithmic phase of growth. DNA quality control was performed by measuring the absorbance at 260/230, concentration was determined using Qubit, and DNA integrity was analyzed by 0.8% agarose gel electrophoresis and using an Agilent 2200 Tapestation system with the Genomic DNA Screen Tape assay. Genome sequencing was performed using Illumina TruSeq DNA PCR-Free kit on Illumina HiSeq 4000 platform with 2 × 100 paired-end reads. Sequence quality metrics were assessed using FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/).

The long reads were obtained using the ONT MinION sequencer on FLO-MIN106 R9v flow cells. We prepared the library using the 1D Genomic DNA by ligation (SQK-LSK108) protocol. Briefly, high molecular weight DNA (1.3 μg) was sheared with a g-TUBE (Covaris) to an average fragment length of 8 Kb. The sheared DNA was repaired using the FFPE Repair mix (New England Biolabs), polished and an A overhang was added with NEBNext End Prep Module (New England Biolabs). Subsequently, adapters (Adapter Mix AMX1D) were ligated using the Blunt/TA Ligase Master Mix (New England Biolabs). Between each step, DNA was cleaned using Ampure XP beads (Beckman Coulter) in a 1:1 proportion. The final library was loaded on the MinION flow cell and monitored by MinKNOW software (version 1.15.1) during a 48-h sequencing time. Generated reads were base-
called in real time and assembled using Canu v1.4 [13]. The assembly was corrected using PILON [14]. The final assembled version was assessed by QUAST (Quality Assessment Tool for Genome Assemblies) [15] in Icarus genome browser [16]. The Companion webtool (https://companion.sanger.ac.uk/) was used for gene prediction and annotation, and Leishmania major as a reference genome [17].


**Funding**: This research was funded in part by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001, CAPES-COFECUB (Comité Français d’Évaluation de la Coopération Universitaire et Scientifique avec le Brésil, program 923/18), European Regional Funds (project “Centre for Research of Pathogenicity and Virulence of Parasites” CZ.02.1.01/0.0/0.0/16_019/0000759), Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Fundação de Amparo a Pesquisa do Estado do Rio de Janeiro (FAPERJ) and Fundação Oswaldo Cruz (Fiocruz). P.B. is a Scientific Logistics Manager of the Genomics Platform of University of Louvain. C.M.dL. was awarded a scholarship from the CAPES Foundation (PVEX - 88881.169912/2018-01), in order to conduct part of his Research as a Visiting scholar at Université de Louvain (Brussels/Belgium).

**Acknowledgments**: The correspondence author is particularly grateful for COBIO from CNPq, the scientific and career experience were deeply boosted by the extra challenges granted by this funding agency. The authors thank the National Lottery (Belgium) and the Foundation against Cancer (2010-101, Belgium) for their support to the Genomics Platform of University of Louvain and de Duve Institute, as well as the Fonds de la Recherche Scientifique - FNRS Equipment Grant UN035.17 for the «Big data analysis cluster for NGS at UCL».

**Conflicts of Interest**: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

**Appendix A**

**Data access**

This Whole Genome Shotgun project has been deposited at DDBJ/ENA/GenBank under the accession VFSE00000000. The version described in this paper is version VFSE01000000.

**References**


