Draft genome sequence of the anaerobic intestinal parasite Blastocystis subtype 4 (ST4)

Ivan Wawrzyniak, Damien Courtine, Céline Nourrisson, Philippe Poirier, Amandine Cian, Aldert Bart, Magali Chabé, Léa Siegwald, Valérie Polonais, Abdel Belkorchia, et al.

Draft genome sequence of the anaerobic intestinal parasite *Blastocystis* subtype 4 (ST4)

Ivan Wawrzyniak [1], Damien Courtine [1], Céline Nourrisson [1], Philippe Poirier [1], Amandine Cian [2], Aldert Bart [3], Magali Chabé [2], Léa Siegwald [4], Valérie Polonais [1], Abdel Belkorchia [1], Tom Van Gool [3], Eric Viscogliosi [2], Frédéric Delbac [1]*

1 : Laboratoire Microorganismes : Génome et Environnement (LMGE), UMR 6023 CNRS-Université Blaise Pascal -Université d'Auvergne, Campus des Cézeaux, 24, avenue des Landais, BP 80026, 63 170 AUBIERE - France
2 : Center for Infection and Immunity of Lille (CIIL), Institut Pasteur de Lille, Inserm U1019, CNRS UMR 8204, Université Lille Nord de France, Biology and Diversity of Emerging Eukaryotic Pathogens, F-59019 Lille Cedex-France
3 : Center for Infection and Immunity Amsterdam (CINIMA), Parasitology Section, Department of Medical Microbiology, Academic Medical Center, Amsterdam - Pays-Bas
4 : Genoscreen Campus Pasteur, 59000 Lille, France - France
* : Corresponding author

Introduction : *Blastocystis* is a highly prevalent protozoa of the intestinal tract belonging to the Stramenopile group. Its prevalence in human often exceeds 5% in industrialized countries reaching as high as 76% in developing countries. Although the role of *Blastocystis* as human pathogen remains unclear, it can cause acute or chronic digestive disorders and some studies have suggested an association with irritable bowel syndrome. The life cycle of the parasite is poorly documented, the vacuolar stage being the most easily recognizable and the most frequently observed in both laboratory culture and stool samples (Fig. 1). *Blastocystis* exhibits an extensive genetic diversity. Seventeen subtypes (ST1-ST17), among which the first nine are found in human, have been identified based on the gene coding for the small-subunit ribosomal RNA.
We have previously provided the first genome sequence of a *Blastocystis* ST7 isolate (Denoeud *et al.,* 2011). It consists of a 18.8 Mb nuclear genome with 6020 genes and a circular DNA molecule of 29 Kb located within mitochondria-like organelles (MLO).

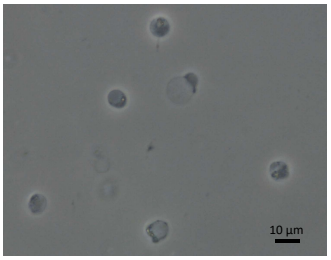

Fig 1. *In vitro* cultivated vacuolar forms of *Blastocystis* ST4 by phase-contrast microscopy

Here we report the sequencing and annotation of the genome of a *Blastocystis* ST4 isolate. The genome sequencing was done with illumina HiSeq 2000 system. The reads were de novo assembled using the IDBA-ud algorithm. Genes were predicted using the Maker annotation pipeline. We also used this pipeline to improve the annotation of the *Blastocystis* ST7 genome

Sequencing and assembly of the ST4 genome

ST4 genome sequencing generated more than 43 millions of 100-bp paired-end reads. After assembly, 3996 scaffolds higher than 200 bp were obtained. Assembly also provided a circular genome of 27 kb in size corresponding to the whole MLO genome

	scaffolds	Scaffolds N50 kb	size (Mb)	G+C %	Number of genes	MLO genome size bp	G+C %	Number of genes
***Blastocystis* ST4**	3996	20,4	13,36	39,7	6046	27815	21,94	45
***Blastocystis* ST7**	54	900,6	18,8	45,2	7098	29270	20,03	45

Maker annotation pipeline

The MAKER genome annotation pipeline was used to annotate the ST4 genome and improve the annotation of the ST7. Using this pipeline, 3656 and 6298 gene models were predicted for ST4 and ST7 respectively. 3665 and 2350 without any homology (*ab initio* genes) were also predicted for the ST4 and ST7.

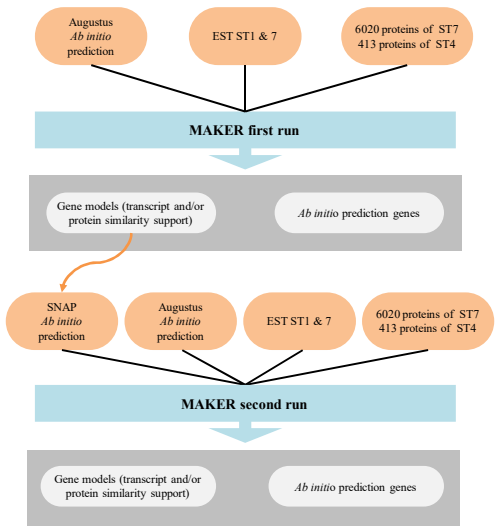

Fig 2. Schematic representation of the MAKER annotation pipeline used for gene prediction. The first run was performed using the *ab initio* predictor Augustus trained with 413 genes manually designed from the ST4 scaffolds, the 6020 annotated genes of the ST7 genome and available ESTs data from both ST7 and ST1. Gene models obtained from the first run were then used to train another *ab initio* gene prediction program called SNAP. A second run of Maker similar to the first run and including the newly trained gene predictor SNAP was finally performed.

Final annotation

For the final annotation, all the gene models were conserved. The *ab initio* predicted genes were first filtered for the presence of InterPro protein domain. All the genes with no domain were secondly filtered by a BLASTP against nr database. We only conserved genes with an e-value lower than 10^{-5}. This led to a final annotation set of 6046 genes for ST4 and 7098 genes for ST7. Gene functions were then annotated by Blast2GO and BLASTP analyses with NCBI, Swissprot/Uniprot and KEGG databases. OrthoMCL was applied to compare both ST4 and ST7 genomes

Comparison with other Stramenopile parasites

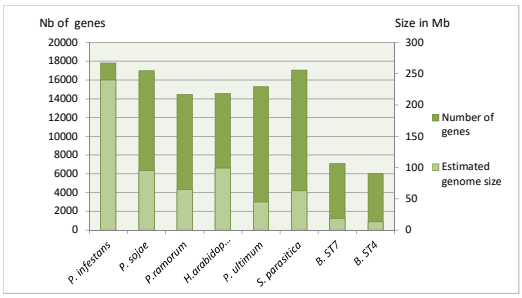

Fig 3. Genome size and number of genes in different Stramenopile parasites. *Blastocystis* is characterized by a smaller genome size and a reduced number of genes in comparison with other stramenopiles

Comparison of *Blastocystis* ST4 and ST7 genomes

To identify proteins specific to each *Blastocystis* ST *vs* conserved proteins between ST4 and ST7 (« core proteome »), orthologs and close paralogs were clustered using OrthoMCL algorithm (Fig 4).

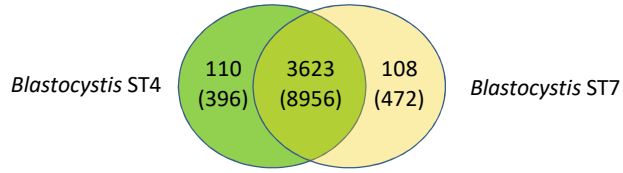

Fig 4. Gene families shared by *Blastocystis*. Of the 13144 protein-coding genes (6048 of ST4 + 7098 of ST7), 9824 genes clustered into 3841 gene families. A total of 8956 genes that clustered into 3623 gene families were common to both ST4 and ST7. A total of 110 families containing 396 genes were specific to ST4 whereas a total of 108 families containing 472 genes were specific to ST7. 1715 and 1605 genes were singleton in ST4 and ST7, respectively (not shown in the figure).