Analysis of Expressed Sequence Tags (ESTs) from the entomopathogenic alga *Helicosporidium* sp. (Chlorophyta, Trebouxiophyceae)

Aurélien Tartar[1], Audrey P. de Koning[2], Patrick J. Keeling[2] and Drion G. Boucias[1]

[1] Department of Entomology & Nematology, University of Florida, Gainesville, USA
[2] Department of Botany, University of British Columbia, Vancouver, Canada

SUMMARY

- **A total of 1288 Expressed Sequence Tags (ESTs) have been generated from a *Helicosporidium* sp. cDNA library.**
- **Only half (48.5 %) of these clones exhibited a significant similarity with a sequence listed in the NCBI database.**
- **A unigene set of 366 identifiable contigs was established from the 624 sequences that exhibited significant similarity.**
- **The phylogenetic signal of the majority of these contigs confirmed *Helicosporidium* sp. as a member of the green (Plants and Algae) lineage.**
- **Several clones were found to be similar to plastid-targeted genes, suggesting that *Helicosporidium* sp. has retained a chloroplast-like organelle.**

INTRODUCTION

The Helicosporidia are obscure pathogenic protists that have been reported to infect a wide range of invertebrate hosts. They have been studied so little that their occurrence and importance as invertebrate pathogens are unclear. Only one species, *Helicosporidium parasiticum*, has been described. Following the recent isolation of a new *Helicosporidium* sp. in Florida, we compiled morphological (1) and molecular data (2, 3) and demonstrated that the Helicosporidia are non-photosynthetic green algae, and that they are related to *Prototheca*, another non-photosynthetic, parasitic algal genus (Chlorophyta, Trebouxiophyceae). Several independent phylogenetic analyses showed that *Helicosporidium* sp. clusters within the class Trebouxiophyceae, in a monophyletic clade that contains *Prototheca* spp. and *Auxenochlorella protothecoides*, suggesting that these organisms arose from a common ancestor (2, 3, 4). The Helicosporidia are unique organisms, which remain poorly characterized at the molecular level. Significantly, they represent to date the only known entomopathogenic algae. As such, they likely have evolved distinct genetic and biochemical mechanisms for interacting with invertebrates. In an effort to better characterize the biology of *Helicosporidium*, we have initiated a large–scale sequencing project, generating Expressed Sequence Tags (ESTs) from a *Helicosporidium* sp. cDNA library.

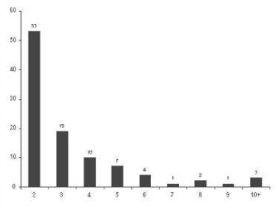

FIG 1: EST redundancy in contig assembly. Most unigenes (266 out of 366) were represented only once. The remaining 100 contigs contained 2+ copies.

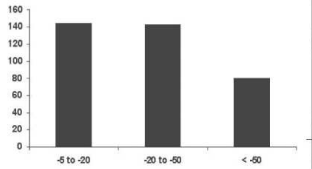

FIG 2: Sequence similarities between *Helicosporidium* sp. ESTs and the best match in the NCBI database. The frequency of the resulting E-value is shown.

MATERIALS AND METHODS

The *Helicosporidium* sp. isolate was maintained on artificial media (TC insect medium supplemented by Fetal Calf Serum) and incubated at 26 ˚C (1). Cells were collected by low speed centrifugation, resuspended into 10 ml of TriReagent (Sigma) plus glass beads (0.45 mm), and broken using a Braun MSK homogenizer. Following cell breakage, total RNA was extracted using the TriReagent manufacturer protocol. Poly(A) mRNA was isolated using the Oligotex mRNA purification kit (Qiagen) and stored at –70 ˚C. The cDNA library was prepared in the Uni-ZAP XR plasmid using the ZAP-cDNA synthesis kit (Stratagene). Following the mass excision protocol, *E. coli* colonies were controlled by PCR for presence of an insert and transferred to 96-well plates. Plates were processed for sequencing both at the University of Florida and University of British Columbia. Sequencing of the cDNA clones was performed from the 5´ end using the T3 primer. Automated sequence similarity searches were done for each ESTs using the BlastN and BlastX algorithms to identify putative gene homologues in the non redundant protein sequence database of the NCBI. BlastX E-values were used as a measure of sequence similarity, and ESTs with E-values < 10^{-5} were assigned to functional classes based on the functional catalog of plant genes. Selected sequences were aligned with representative eukaryotic homologues in ClustalX and phylogenetic relationships were calculated in PAUP* using parsimony and distance methods. Branch support was assessed by bootstrapping (100 replicates).

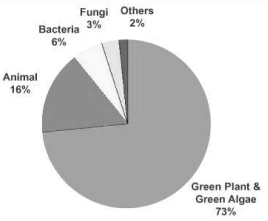

FIG 3: Taxonomic distribution of the closest homologue of *Helicosporidium* sp. unigenes. The 366 unigenes with significant similarity were classified according to the species with the best BlastX scores for each unigene (E-value lower than 10^{-5}). Green plants and green algae accounted for most (73%) of the matches.

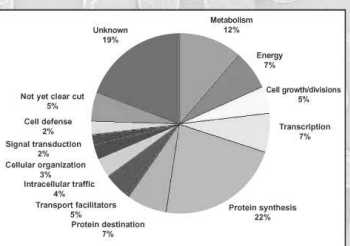

FIG 4: Functional classification of *Helicosporidium* sp. ESTs. The 366 unigenes were classified according to their putative function (determined by similarity).

EST DATABASE ANALYSIS

A total of 1288 clones were generated by random sequencing of a cDNA library from *Helicosporidium* sp. Similarity searches showed that half of these sequences do not possess any homologues in the NBCI non-redundant database, indicating that those clones may represent *Helicosporidium*-specific genes. Thus, it suggests that the potential for new gene discovery in the *Helicosporidium* transcriptome is very high. The other half corresponds to 624 sequences with significant similarity to known sequences (E-values lower than 10^{-5}). A set of 366 contigs was assembled from these sequences (Fig 1) and further analyzed. A high proportion of these contigs was shown to have very significant similarity to known protein sequences, with an E-value lower than 10^{-20} (Fig 2). These high similarity values allowed for the assignment of both a closely related species and a putative function for each unigene. Therefore, the 366 unigenes were then classified according to the taxonomic distribution of their closest homologues (Fig 3), and according to their functional categories (Fig 4). These categories have been determined following the functional catalog of plant genes established for the analysis of the *Arabidopsis thaliana* genome (5). Significantly, 24% of the contigs are similar to protein sequences for which the function remains unclear or unknown, thereby lowering even more the final number of truly identifiable genes: 277 genes were identified with confidence, out of our 1288 sequenced clone effort. This small number of unigenes highlights the uniqueness of *Helicosporidium* sp. yet is sufficient to provide insights into its biology.

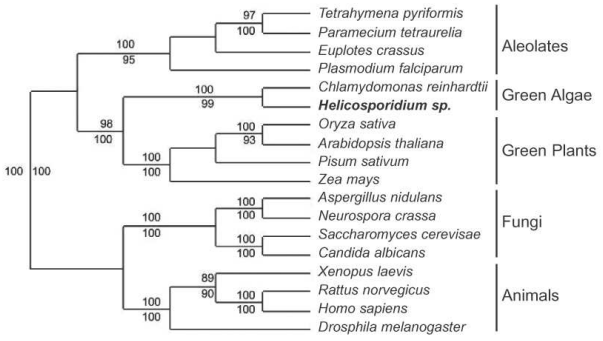

FIG 5: Phylogenetic (Neighbor-Joining) tree based on a concatenated dataset (1235 characters) containing four protein sequences corresponding to the actin, alpha tubulin, beta tubulin and glyceraldehyde 3 phosphate dehydrogenase (GAPDH) genes. The alpha tubulin (consensus sequence of clones 12G01 and 14A09) and GAPDH (clone 5F07) sequences from *Helicosporidium* sp. were obtained from the EST, whereas the actin and beta tubulin fragments were sequenced from PCR-amplification of genomic DNA. The tree depicts *Helicosporidium* sp. as a green alga (Chlorophyta) and this relationship is highly supported by bootstrap values.

SELECTED SEQUENCES ANALYSIS

Two contigs (alpha tubulin and GAPDH) were selected to be used in phylogenetic analyses (Fig 5) because they are highly conserved and a wide variety of eukaryotic homologues are available in public databases. The major eukaryotic lineages are represented and their monophylies and relationships to each other are very strongly supported by bootstrap values. This analysis confirmed that Helicosporidia are non-photosynthetic green algae.

Additionally, several plastid-targeted genes have been characterized by similarity with plant and algal genes. This suggests that the Helicosporidia, despite having lost their photosynthetic ability, have retained a chloroplast-like organelle. Some of the clones were found to exhibit a 5´ leader sequence that is consistent with plastid targeting (6). These genes have been associated with several metabolic pathways, leading to the hypothesis that the Helicosporidia plastid may be used for amino acid and fatty acid biosynthesis.

Metabolic pathway	gene discovered in ESTs
Gene expression	rpl19, rpl15, polyadenylate-binding protein, ribonucleoprotein M
Protein import	ClpB, Cpn10
Cysteine biosynthesis	cysteine synthase, adenylysulfate kinase
Leucine biosynthesis	2-isopropylmalate synthase, 3-isopropylmalate dehydratase
Aspartate biosynthesis	asparate aminotransferase
Serine biosynthesis	phosphoserine aminotransferase
Type II fatty acid biosynthesis	acyl carrier protein, FabG
Fatty acid modification	ACP-stearoyl desaturase
Isoprenoid biosynthesis	LytB
CO2 fixation	carbonic anhydrase
Tetrapyrrole biosynthesis	glutamate-1-semialdehyde 2, 1-aminomutase
Other nucleotide	ferredoxin, ferredoxin-thioredoxin reductase, ycf24, diphosphate kinase, PDHC E3 dihydro-

Table 1: List of all the *Helicosporidium* sp. nuclear-encoded, plastid-targeted genes identified in the EST database. The genes are classified according to their putative function.

REFERENCES

1. **Boucias, DG, Becnel, JJ, White, SE & Botts, M.** (2001) J. Eukaryot. Microbiol. 48, 460-470
2. **Tartar, A, Boucias, DG, Adams, BJ & Becnel, JJ** (2002) Int. J. Syst. Evol. Microbiol. 52, 273-279
3. **Tartar, A, Boucias, DG, Becnel, JJ & Adams, BJ** (2003) Int. J. Syst. Evol. Microbiol. In press
4. **Ueno, R, Urano, N, Suzuki, M** (2003) FEMS Microbiol. Letters 223, 275-280
5. **Bevan, M, et al.** (1998) Nature 391, 485-488
6. **Waller, RF, et al.** (1998) Proc. Natl. Acad. Sci. USA 95, 12352-12357

UNIVERSITY OF FLORIDA