Algal genomes reveal evolutionary mosaicism and the fate of nucleomorphs

Bruce A. Curtis1,2,3, Goro Tanifuji1,2,3, Fabien Burki1,2,3, Ansgar Gruber4,5, Manuel Irimia6, Shinichiro Maruyama1,2,3, Maria C. Arias7, Steven G. Ball8, Gillian H. Liu1,2,3, Yoshihisa Hirakawa4,5, Julia F. Hopkins1,2,3, Alan Kuo8, Stefan A. Rensing4, Jeremy Schmutz6,10, Alkaterei Symbioinid9, Marek Elias1,2, Robert J. M. Eveleigh1,2,11, Emily K. Herman13, Mary J. Klute13, Takuro Nakayama2,3, Miroslav Oborník14,15,16, Adrian Reyes-Prieto1,17, E. Virginia Armbrust18, Stephen J. Aves19, Robert G. Beiko20, Pedro Coutinho21, Joel D. Durford17, Naomi M. Fast4, Beverley R. Green4, Cameron J. Grisdale4, Franziska Hempel22, Bernard Henrisat21, Marc P. Hoppner23, Ken-Ichi Ishida24, Eunsoo Kim25, Luděk Kofín14,15, Peter G. Kroth3, Yuan Liu9,26, Shehe Rai-Banoo Malik1,2,3, Uwe G. Maier22, Darcy McRose27, Thomas Mock28, Christopher E. Lane1,2,3,5, Patrick J. Keeling3,4, Michael W. Gray1,2,3,35, Igor V. Grigoriev8 & John M. Archibald1,2,3,35

The photosynthetic organelles (plastids) of algae evolved from cyanobacteria by endosymbiosis1,2. The ‘primary’ plastids of red algae, glucophyte algae and green algae, and their land-plant descendants, probably arose just once, more than a billion years ago3,4. Subsequent to this key event, the primary plastids of red and green algae were laterally transferred to other eukaryotes by secondary and tertiary endosymbioses, spanning some of the most abundant and ecologically important aquatic photosynthesizers on Earth such as diatoms, giant kelp, bloom-forming haptophytes and toxic dinoflagellates, as well as parasites such as the malaria pathogen Plasmodium5.

We have sequenced the nuclear genomes of two unicellular algae that are remarkable in their genetic and cellular complexity: the cryptophyte Guillardia theta and the chlorarachniophyte Bigelowiella natans. The secondary plastids of these independently evolved algae are unique in retaining a relict endosymbiont nucleus (the nucleomorph). Cryptophyte and chlorarachniophyte cells thus have four genomes and require complex subcellular protein-targeting machinery and inter-compartment coordination (Fig. 1). The B. natans nuclear genome is the first to be sequenced from a rhizarian protist, and the G. theta nuclear genome sequence is the first from a cryptophyte. They

©2012 Macmillan Publishers Limited. All rights reserved
The nuclear genomes of B. natans and G. theta are essentially ‘complete’ with respect to the major hallmarks of eukaryotic cellular complexity (Supplementary Information 2.6), the relict nucleus of the eukaryotic endosymbiont. The predicted numbers of Nuclear and plastid genes are rich (>21,000 predicted protein genes each, >85% of which are supported by RNA-seq data). Of the inferred proteins, 51% in G. theta and 47% in B. natans are unique, that is, have no detectable homologues in any other organism. Both genomes contain a large number of paralogues, constituting 2,636 multi-gene families in B. natans and 3,284 in G. theta (Supplementary Table 1.6.2).

As inferred from functional classifications based on the euKaryotic Orthologous Groups (KOOG) database, and protein family analyses (Supplementary Information 2.6), the G. theta and B. natans genomes are essentially ‘complete’ with respect to the major hallmarks of eukaryotic cellular complexity (>97% of a set of ‘core eukaryotic genes’ is present in both organisms). These include components of the endomembrane system (Supplementary Information 2.6.3), transcription, RNA processing and modification, post-translational modification and splicing at 3' splicing of alternatively spliced isoforms (Supplementary Information 2.2). Never-theless, even if most of the splicing complexity seen in B. natans simply reflects mis-splicing, many of these alternative transcripts might have important functions. A systematic survey of RNA-seq data identified 246 cases of genes whose alternative isoforms differentially include or exclude amino-terminal signal-peptide-encoding regions (three of which were verified by RT–PCR), suggesting that alternative splicing has a role in the generation of proteins targeted to different subcellular compartments (below). Alternative splicing has recently been shown to mediate dual targeting of glycolytic enzymes to the cytosol and peroxisome in fungi.

**Subcellular proteomes**

Cryptophyte and chlorarachniophyte nucleomorphs are residual, endosymbiotic nuclei with tiny genomes <1 Mb in size4-17. The G. theta and B. natans nucleomorph genomes have only 487 (ref. 17) and 331 (ref. 15) protein genes, respectively, comprised of a limited set of ‘housekeeping’ genes, 31 or fewer genes for plastid-targeted proteins, and an abundance of ‘ORFan’ genes that typically show no detectable sequence similarity to known proteins41.
Like plastids and mitochondria, nucleomorphs and their genomes are reduced beyond self-sufficiency; they depend on nucleus-encoded proteins that are targeted to the periplastidial compartment (PPC), the residual endosymbiont cytoplasm in which the nucleomorph resides (Fig. 1). However, only a handful of PPC-targeted proteins are known (for example, see refs 18–21) and the true extent of this dependence is unclear. Indeed, why nucleomorph genomes have been retained at all is a long-standing mystery of plastid evolution. Bearing in mind our knowledge of the G. theta and B. natans plastid\textsuperscript{22,23} nucleomorph\textsuperscript{15,17} and mitochondrial (this study) genome sequences, we carried out a comprehensive examination of nucleus-encoded proteins predicted to be targeted to each subcellular compartment (Supplementary Information 1.9), with emphasis on the PPC.

Our in silico-predicted mitochondrial, plastid, and PPC and nucleomorph proteomes for G. theta and B. natans are summarized in Supplementary Table 1.9.1 and Supplementary Fig. 1.9.4.1.2, together with a predicted set of >600 evolutionarily conserved endoplasmic reticulum and Golgi proteins (Supplementary Information 1.9). The limited overlap in proteins predicted to be targeted to different compartments suggests that the search strategies successfully differentiated among plastid-, PPC- and nucleomorph-, and host endoplasmic-reticulum- and Golgi-targeted proteins, which is important because in both cryptophytes and chlorarachniophytes the signal peptide secretion system is the first step in trafficking proteins to each of these compartments\textsuperscript{1}. We analysed these proteomes in order to compare and contrast the biology of the independently evolved plastid and periplastidal compartments in G. theta and B. natans.

G. theta is predicted to have twice as many PPC- and nucleomorph-targeted proteins as B. natans (2,401 versus 1,002, after removal of ambiguously assigned proteins). A KOG-based breakdown of the unique (that is, non-overlapping) proteins in the PPC or nucleomorph proteomes revealed three KOG categories that are particularly ‘enriched’ in G. theta relative to B. natans: post-translational modification, protein turnover and chaperones; signal transduction; and carbohydrate transport and metabolism (Fig. 2a and Supplementary Table 1.9.4.1.1). The biological significance of these observations was further revealed through the mapping of nucleomorph- and nucleus-encoded, PPC-targeted proteins to KEGG (Kyoto Encyclopedia of Genes and Genomes) metabolic pathways (Supplementary Information 2.4). The G. theta PPC possesses canonical components of the protein-degrading proteasome, which B. natans seems to have lost entirely (KEGG pathway map 03050, Supplementary Fig. 2.4.1). G. theta also seems to have more PPC-localized proteins dedicated to protein folding (molecular chaperones) and RNA degradation, and a much greater diversity of metabolic enzymes, including those involved in amino acid biosynthesis (Supplementary Information 2.6.1 and Supplementary Fig. 2.4.1). In contrast, B. natans has a larger number of predicted nucleomorph-localized, spliceosome-associated proteins than does G. theta (Supplementary Fig. 2.4.1), which correlates with the marked difference in intron abundance: 852 in the B. natans nucleomorph genome\textsuperscript{15} versus just 17 in G. theta\textsuperscript{22}.

Host nuclear control over organelle biology is apparent in both G. theta and B. natans in the form of nucleus-encoded transcription-associated proteins (presumably regulat in G. theta gene expression; Supplementary Information 2.6.1), putative DNA replication machinery, and ‘cell cycle-related’ proteins (for example, protein kinases) (Supplementary Fig. 2.4.1). Other processes in the PPC and nucleomorph are driven mainly by nucleomorph-encoded proteins, translation being a prominent example (Fig. 2a; KOG category J (translation, ribosomal structure and biogenesis)). Near-complete repertoires of small and large PPC ribosomal subunits could be inferred for G. theta, somewhat less so for B. natans, and in both cases the bulk of the constituent proteins are nucleomorph-encoded (Supplementary Fig. 2.4.2). This mirrors the pattern seen in plastid genomes\textsuperscript{26}, in which core processes such as transcription and translation primarily involve proteins synthesized ‘on-site’.

Carbon metabolism differs substantially in G. theta and B. natans as inferred from the identification of putative carbohydrate-active enzymes (Supplementary Information 1.11 and Tables 2.3.1 and 2.3.2). Subcellular mapping of putative glycolysis-associated proteins in G. theta (Supplementary Fig. 2.4.5) reveals many PPC-localized reactions catalysed by key enzymes such as glucan, water dikinase, alpha-amylase, hexokinase, 6-phosphofructokinase and phosphoglucomutase. These enzymes form a link to the synthesis and degradation of starch, which occurs in the PPC in G. theta\textsuperscript{22}. Thirty-six candidate proteins for PPC, plastid or cytosol metabolite shutting

---

**Figure 2** | Complexity of the periplastidial compartment in cryptophytes and chlorarachniophytes. a, Histogram showing the number of proteins predicted to be targeted to the PPC of G. theta and B. natans broken down by KOG functional category. For each KOG category, nucleomorph (NM)- and nucleus (NU)-encoded proteins are shown (PPC proteins predicted to be targeted to more than one subcellular compartment were removed; see Supplementary Fig. 1.9.4.1.2). b, Histogram showing the diversity of protein functions in the G. theta and B. natans PPC relative to free-living organisms (colour-coding as in a). Numbers of distinct KOG identifiers (IDs) in the PPC proteomes are plotted as a percentage of the average number of KOG IDs across 25 KOG categories for 6 organisms: Chlamydomonas reinhardtii, Ostreococcus tauri, Arabidopsis lyrata, Emiliania huxleyi, Dictyostelium purpureum and Phaeodactylum tricornutum (see Supplementary Information 1.9.3). Plastid and mitochondrial proteins were removed before calculating the averages (see Supplementary Information). KOG categories are as follows: A, RNA processing and modification; B, chromatin structure and dynamics; C, energy production and conversion; D, cell cycle control, cell division and chromosome partitioning; E, amino acid transport and metabolism; F, nucleotide transport and metabolism; G, carbohydrate transport and metabolism; H, coenzyme transport and metabolism; I, lipid transport and metabolism; J, translation, ribosomal structure and biogenesis; K, transcription; L, replication, recombination and repair; M, cell wall, membrane or envelope biogenesis; N, cell motility; O, post-translational modification, protein turnover, chaperones; P, inorganic ion transport and metabolism; Q, secondary metabolites biosynthesis, transport and catabolism; R, general function prediction only; S, function unknown; T, signal transduction; U, intracellular trafficking, secretion and vesicular transport; V, defence mechanisms; W, extracellular structures; Y, nuclear structure; Z, cytoskeleton. Higher KOG categories are as follows: CP, cellular processing and signalling; Hyp, poorly characterized; Inf, information storage and processing; Met, metabolism.
in *G. theta* were identified from a set of 757 putative membrane-transport-associated proteins (Supplementary Information 1.10 and Supplementary Table 1.10.2). The distribution of glycolytic enzymes in *B. natans* is very different from that of *G. theta*, with a more heterogeneous mix of PPC-, plastid-, mitochondrion- and host cytosol-localized proteins (Supplementary Figs 2.4.3 and 2.4.4). In chlorarachniophytes the main carbohydrate storage product is a β-1,3-glucan located in the host cytoplasm26 (Fig. 1), and we identified numerous enzymes that are likely to have roles in β-glucan metabolism (Supplementary Information 2.3.2.1 and Supplementary Table 2.3.4).

We next examined the reduction in the PPC and nucleomorph proteomes of cryptophytes and chlorarachniophytes relative to the free-living organisms from which they evolved. We used the number of different KOG identifiers present in each of the 25 KOG functional categories as a measure of the diversity of biochemical processes taking place in the *B. natans* and *G. theta* PPC (taking into account nucleomorph-encoded proteins) (Supplementary Information 1.9.4.3). A total of 237 and 452 unique KOG identifiers were assigned to the *B. natans* and *G. theta* PPC proteome data sets, respectively (Supplementary Table 1.9.4.3.1). For most KOG categories the number of KOG identifiers in *G. theta* and *B. natans* is <25% of the average calculated from a set of 6 free-living organisms (algae with primary and secondary plastids plus a heterotrophic amoeba; Fig. 2b and Supplementary Information 1.9.4.3). Some functional categories are, predictably, completely absent in both organisms (for example, *n* = cell motility), whereas in *G. theta* the number of KOG identifiers in three different KOG categories exceeds 40% of the ‘free-living’ average (category I (translation, ribosomal structure and biogenesis), G (carbohydrate transport and metabolism) and P (inorganic ion transport and metabolism); Fig. 2b). On balance, the PPC of cryptophytes and chlorarachniophytes is highly reduced, but has retained an unexpectedly broad range of biochemical processes. These data provide the basis for addressing many fundamental questions about algal cell biology, including how many homologues of *G. theta* and *B. natans* PPC proteins are retained in the nucleomorph-lacking PPC of algae such as diatoms and haptophytes25, and what exactly are the biochemical determinants of the protein trafficking pathways in cryptophytes, chlorarachniophytes and other secondary plastid-bearing algae26–27. Making sense of the hundreds of predicted PPC and nucleomorph proteins in *G. theta* and *B. natans* with unknown functions (Supplementary Table 1.9.4.1.1) will be a substantial challenge.

Endosymbiotic gene transfer and replacement

Endosymbiotic gene transfer (EGT)—the movement of DNA from endosymbiont to host before, during and after the evolution of an organelle—has had a notable role in the evolution of algae and their nuclear genomes28–29. The genomes of eukaryotes that are known or thought to have once had red algal endosymbionts are now regularly queried for the presence of endosymbiotic genes (for example, refs 30, 31). Organisms with (or thought to have once had) red algal secondary plastids are predicted to have ‘red’ genes in their nuclear genomes and ‘green’ genes should be found in the nuclei of organisms with green algal secondary plastids. Quantification of these algal signatures has the potential to answer fundamental questions about the spread and secondary loss of plastids across the eukaryotic tree but has led to conflicting results. For example, a large and unexpected number of ‘green’ genes were found in the genomes of diatoms32 and were interpreted to be evidence of a cryptic secondary endosymbiosis involving a green alga before the establishment of the red alga-derived plastid that diatoms currently harbour. However, these results were re-evaluated and found to be unconvincing33. The ability to detect genes of algal origin in nuclear genomes and to accurately distinguish between ‘red’ and ‘green’ has been shown to be complicated by a number of factors including taxonomic sampling bias, phylogenetic artefacts and large data sets consisting of thousands of complex trees that are invariably processed in an automated fashion31,33,34. We carried out a comprehensive phylogenomic investigation of EGT in the nuclear genomes of *B. natans* and *G. theta*, whose plastids and nucleomorphs are of green and red algal ancestry, respectively, using protocols and programs designed to address, to the extent possible, those issues mentioned above and other potential problems (Supplementary Information 1.12 and Supplementary Fig. 1.12.3).

From a set of 6,181 *B. natans* genes for which protein-based phylogenetic trees could be generated, automated tree sorting and manual curation resulted in the identification of 353 genes (5.7%) for which an algal origin could be confidently inferred (Fig. 3 and Supplementary Fig. 1.12.3). As expected, a large proportion (207; 59%) of these were green algal in nature, although 45 (22%) were classified as being derived from red algae. This pattern resembles that seen in an early EST-based analysis of *B. natans* proteins and was attributed to the mixotrophic lifestyle of chlorarachniophyte algae35. For *G. theta*, 508 of 7,451 genes (6.8%) were deemed to be algal in origin (Fig. 3 and Supplementary Fig. 1.12.3). Interestingly, more than twice as many *G. theta* genes in the manually curated set were classified as green (252) than red (100), and 9 examples of apparently glucophyte-derived genes were identified. These results should be interpreted with caution, however, because although our analyses included all available red algal protein data sets (Supplementary Table 1.12.1), taxon sampling is still biased towards ‘green’ lineages. In fact, the majority (143 out of 252) of the protein trees for ‘green’ genes in *G. theta* contained no red algal homologues (Fig. 3) and 147 out of 508 were considered ‘algal’ but ambiguous with respect to which type. Thus, increased taxon sampling from red algae will presumably affect several predictions and perhaps enable more meaningful interpretation of others, as underscored by recent authors investigating ‘red’ and ‘green’ signals in other organisms with red secondary plastids31,33,34. The same can be said for interpretation of the red algal genes in *B. natans*, which has a green alga-derived secondary plastid. These uncertainties are exacerbated further by the still-unresolved phylogenetic position of the host component of cryptophytes relative to primary and secondary plastid-bearing algae36. Consequently, testing hypotheses about possible biological explanations for the diversity of algal nuclear genes seen in *G. theta* and *B. natans*, such as the relative contributions of endosymbiotic versus horizontal gene transfer, cannot currently be carried out without careful consideration of taxon sampling and methodological artefacts.

Nevertheless, phylogenomic data taken together with subcellular targeting predictions show that the *B. natans* and *G. theta* nuclear genomes possess a complex mosaic of genes whose evolutionary histories do not reliably predict where their protein products function within the cell. A large portion of the alga-derived proteins identified in both organisms seem to function in their host cytosolic compartments, and clear examples of algal proteins targeted to the mitochondrion, endoplasmic reticulum or Golgi apparatus, plastid, and PPC or nucleomorph were also found (Fig. 3b). These results show that during the course of host–endosymbiont integration proteins often acquire new functions and/or new locations in which to function.

Gene duplication has also played a part in the ‘re-purposing’ of *G. theta* and *B. natans* nuclear genes of both host and endosymbiont ancestry (Supplementary Information 2.5.2 and Supplementary Fig. 2.5.2.2). Of the 508 ‘algae’ genes in the *G. theta* nuclear genome, 71 were found to belong to paralogous gene clusters (that is, genes that have duplicated subsequent to EGT); in ~25% of these cases, paralogues encode proteins predicted to be targeted to multiple compartments, most often the PPC and host cytosol (Supplementary Fig. 2.5.2.1). A similar picture is seen in *B. natans*. In other cases the opposite pattern is observed, that is, duplication of apparent host-derived genes followed by organelle targeting of one or more
of the protein products, sometimes as compensation for the loss of a nucleomorph gene (below). The present-day composition of the G. theta and B. natans subcellular proteomes is the product of extensive mixing and matching of proteins derived from their hosts and from endosymbionts that have become organelles. No clear pattern in the fates of individual endosymbiont-derived genes (loss, retention, duplication or re-purposing) is apparent.

Why do nucleomorphs persist?

Nuclear mitochondrial DNAs (NUMTs) and/or nuclear plastid DNAs (NUPTs) have been found in most eukaryotes studied so far; rates of EGT seem to vary substantially from lineage to lineage. Although most such transfers involve small, apparently random fragments of organellar DNA that have no notable impact on the nuclear genome, entire genes can be transferred and expressed in their new environment. One explanation for the absence of NUNMs and NUMTs in the B. natans and G. theta genomes is the ‘limited transfer window hypothesis’, which posits that cells with multiple copies of an organelle are more likely to have EGTS than those with single organelles because lysis of a single organelle to release DNA into the host nucleocyttoplasm would be catastrophic. Consistent with our observations, G. theta and B. natans cells have a single plastid–nucleomorph complex per cell, the lysis of which would presumably be fatal. In contrast, cryptophytes can have large, reticulate mitochondria that undergo fission and fusion, and in chlorarachniophytes each cell generally has multiple mitochondria that reside in both the cytoplasm and (when present) filopodia. The ability of the limited transfer window hypothesis to explain the absence of NUPTs and NUNMs in G. theta and B. natans could be tested further by searching the nucleomorph genomes of cryptophytes and chlorarachniophytes that contain multiple plastid–nucleomorph complexes per cell.

The consequences of the lack of NUNMs and NUPTs in the B. natans and G. theta nuclear genomes are considerable. In the absence of EGT, inactivation and loss of essential plastid and nucleomorph genes cannot be compensated for by the classical gene transfer–protein re-targeting scenario, as occurs in other systems. Our results show that indirect ‘solutions’ have evolved, most notably the duplication and functional reassignment of host-derived nuclear genes. For example, a nucleus-encoded cyclin-dependent kinase regulatory subunit protein (also known as kin(cdc)) predicted to function in the G. theta PPC or nucleomorph is not specifically related to kin(cdc) homologues encoded in the nucleomorph genomes of two other cryptophytes, but instead is a recent duplicate of an apparently host-derived homologue (Supplementary Fig. 2.5.2.2a). A similar pattern is seen in a variety of other nucleus-encoded, PPC-targeted proteins in G. theta (Supplementary Information 2.5.5 and Supplementary Table 2.5.5.1). In B. natans, alternative splicing may serve as an additional mechanism for increasing proteome complexity and compensating for the loss of organellar genes. Over recent evolutionary time scales, nucleomorph genome reduction seems to have slowed markedly for lack of an easy solution to the problem of nucleomorph gene loss.
Extensive EGT has nevertheless occurred in the ancestors of B. natans and G. theta. Some of the protein products of these transferred genes are targeted to the plastid and PPC but most are not (Fig. 3b). Genetic and biochemical mosaicism is thus rampant in both organisms, with host-, endosymbiont- and foreign alga-derived proteins contributing to processes taking place in their various subcellular compartments. The extent to which such mosaicism exists in other cryptophytes and chlorarachniophytes remains unknown. Nevertheless, it seems likely that close inspection of the genomes of all algae that evolved by eukaryote–eukaryote endosymbiosis will reveal a level of mosaicism beyond that which is typically assumed. This conclusion has important implications for the use of genomic data to infer a robust tree of eukaryotes that includes secondary and tertiary plastid-bearing phototrophs, and more generally, for our understanding of the evolution of the eukaryotic cell.

METHODS SUMMARY

DNA was extracted from axenic cultures established from single-cell isolates of *Guillardia theta* and *Bigelowiella natans* (Bigelow Laboratory for Ocean Sciences). Three different-sized libraries, 3-kb, 8-kb and 34-kb fosmids, were generated and sequenced at the Joint Genome Institute (JGI) using Sanger Sciences). Three different-sized libraries, 3-kb, 8-kb and 34-kb fosmids, were generated before the genome projects, JGI-generated ESTs and RNA-seq alternative splicing analyses were assisted by three messenger RNA data sets: ESTs generated at the Joint Genome Institute (JGI) using Sanger Sciences). DNA was extracted from axenic cultures established from single-cell isolates of *Guillardia theta*. Some of the protein products of these trans-ferred genes are targeted to the plastid and PPC but most are not...
50. Martin, W. et al. Gene transfer to the nucleus and the evolution of chloroplasts. Nature 393, 162–165 (1998).

Supplementary Information is available in the online version of the paper.

Acknowledgements The work conducted by the US Department of Energy Joint Genome Institute is supported by the Office of Science of the US Department of Energy under Contract no. DE-AC02-05CH11231. RNA-seq data used in the paper were generated by the National Center for Genome Resources as part of the Gordon and Betty Moore Foundation’s Marine Microbial Eukaryote Transcriptome Project. B.A.C. and J.F.H. were supported by a Special Research Opportunities Grant from the Natural Sciences and Engineering Research Council of Canada awarded to J.M.A. and M.W.G. J.M.A., P.J.K., M.W.G. and C.H.S. are members of the Canadian Institute for Advanced Research, Program in Integrated Microbial Biodiversity. G.I.M. is an Australian Research Council Federation Fellow and a Howard Hughes International Scholar. We thank R. A. Andersen (Bigelow Laboratories) for assistance with single-cell isolations, C. X. Chan for a tree-sorting PERL script, H. Gutierrez for help with SM protein family annotation, and B. Read for permission to analyse the Emiliania huxleyi genome sequenced by the JGI.

Author Contributions Nucleic acid sample preparation: C.E.L., D.F.S. and J.F.H. Genome and transcriptome sequencing and assembly: J.S., J.G., A.K.B., J.A.C., R.K., E.L. and S.L. Genome annotation and/or analysis: B.A.C., G.T., F.B., A.G., M.I., S.M., M.C.A., S.G.B., G.H.G., Y.H., J.F.H., A.K., S.A.R., J.S., A. Symeonidi, M.E., R.J.M.E., E.K.H., M.J.K., T.N., M.O., A.R.-P., E.V.A., S.J.A., R.G.B., P.C., J.B.D., D.G.D., N.M.F., B.R.G., C.J.G., F.H., B.H., M.P.H., K.-U., E.K., L.K., P.G.K., Y.L., S.-B.M., U.G.M., D.M., T.M., J.A.D.N., N.T.O., A.M.P., E.R.J., T.A.R., G.R., S.W.R., C.S., S. Schaack, S. Shirato, C.H.S., S. Suzuki, A.Z.W., S.Z., J.G., A. Salamon, C.E.L., M.W.G. and J.M.A. Project management: K.B., I.V.G. and J.S. Project coordination: J.M.A., M.W.G., P.J.K., C.E.L. and G.I.M. Writing: J.M.A., B.A.C., M.W.G., G.I.M., P.J.K., C.E.L., G.T., F.B., A.G., M.I., S.M., M.C.A., S.G.B., G.H.G., J.F.H., A.K., S.A.R., J.S., A. Symeonidi, R.J.M.E., E.K.H., M.J.K., T.N., A.R.-P., J.B.D.E., E.K., P.G.K., E.J.P., S.W.R., S.S., A.K.B. and I.V.G.

Author Information The G. theta and B. natans genome sequences and annotations are available through the JGI Genome Portal at http://jgi.doe.gov/Gtheta and http://jgi.doe.gov/Bnatans and have been deposited in GenBank under the accession numbers AEIE00000000 and ADNK00000000, respectively. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to J.M.A. (john.archibald@dal.ca).

This work is licensed under a Creative Commons Attribution-NonCommercial-Share Alike 3.0 Unported licence. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-sa/3.0