Perspectives in Science (2015) 6, 84—93

Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/pisc

Euglena in time: Evolution, control of central metabolic processes and multi-domain proteins in carbohydrate and natural product biochemistry☆

Ellis C. O'Neill[a,b], **Martin Trick**[c], **Bernard Henrissat**[d,e], **Robert A. Field**[a,*]

[a] *Department of Biological Chemistry, John Innes Centre, Norwich Research Park, Norwich NR4 7UH, UK*
[b] *Present address: Department of Plant Sciences, University of Oxford, South Parks Road, Oxford, Oxfordshire OX1 3RB, UK*
[c] *Department of Computational and Systems Biology, John Innes Centre, Norwich Research Park, Norwich NR4 7UH, UK*
[d] *Architecture et Fonction des Macromolécules Biologiques, CNRS, Aix-Marseille Université, 13288 Marseille, France*
[e] *Department of Biological Sciences, King Abdulaziz University, Jeddah, Saudi Arabia*

Received 12 March 2015; accepted 28 July 2015
Available online 23 October 2015

KEYWORDS
Euglena gracilis;
Evolution;
Transcriptome;
Natural products;
Base J;
Gene silencing;
O-GlcNAc

Summary *Euglena gracilis* is a eukaryotic microalgae that has been the subject of scientific study for hundreds of years. It has a complex evolutionary history, with traces of at least four endosymbiotic genomes and extensive horizontal gene transfer. Given the importance of *Euglena* in terms of evolutionary cell biology and its unique taxonomic position, we initiated a *de novo* transcriptome sequencing project in order to understand this intriguing organism. By analysing the proteins encoded in this transcriptome, we can identify an extremely complex metabolic capacity, rivalling that of multicellular organisms. Many genes have been acquired from what are now very distantly related species. Herein we consider the biology of *Euglena* in different time frames, from evolution through control of cell biology to metabolic processes associated with carbohydrate and natural products biochemistry.

☆ This article is part of a special issue entitled ''Proceedings of the Beilstein Bozen Symposium 2014 — Chemistry and Time''. Copyright by Beilstein-Institut www.beilstein-institut.de.

* Corresponding author. Tel.: +44 1603 450720; fax: +44 1603 450018.
E-mail address: rob.field@jic.ac.uk (R.A. Field).

http://dx.doi.org/10.1016/j.pisc.2015.07.002

Contents

Introduction

The ease with which *Euglena* can be cultured has made them one of the most highly studied eukaryotes, playing a pivotal role in the development of cell biology and biochemistry. *Euglena gracilis*, in particular, has long been investigated for the production of vitamins A, C, E (Takeyama et al., 1997) and essential amino acids, and is also a good source of polyunsaturated fatty acids (Korn, 1964). When grown aerobically in light it produces an insoluble β-1,3-glucan storage polymer, paramylon (Rodríguez-Zavala et al., 2010), which can make up around 85% of the dry weight of the organism. In contrast, under anaerobic conditions, wax esters comprise over 50% of the dry weight of some strains of *Euglena* (Inui et al., 1982).

Genome sequencing of *Euglena* has been hampered to date due to the large and complex genome (approximately 2 Gbp in size, with 80% repetitive sequence — Mark Field, private communication), which has arisen from a series of endosymbiotic events during its evolution. Aside from typical eukaryotic epigenetic modifications, including DNA methylation and histone acetylation, the genome of *Euglena* also contains the modified nucleotide Base J (glucosylated hydroxythymidine), also found in other kinetoplastids (Borst and Sabatini, 2008), which complicates DNA sequencing by virtue of restricting polymerase processivity. Additionally, *Euglena* has the ability to extensively process mRNA during transcription (Tessier et al., 1992), altering the sequences before translation; hence the proteome of *Euglena* would be difficult to predict from its genome. Avoiding the complications of algal genome sequencing (Rismani-Yazdi et al., 2011) and to begin to explore the full metabolic capability of *Euglena*, we sequenced the transcriptome of *Euglena gracilis* var. *saccharophila* (O'Neill et al., 2015).

Transcript analysis identified 22,814 predicted protein-encoding genes in phototrophic *Euglena* cells, whilst 26,738 were evident in heterotrophic cells, accounting for 32,128 non-redundant predicted proteins overall, including 8890 splice variants. This indicates that there is a dramatic shift in metabolic capability which is dependent upon growth conditions. All of the genes necessary for cellular housekeeping activities are encoded, as well as for the biosynthesis of vitamins, amino acids and complex carbohydrates; a number of novel protein sequences are evident whose activity is difficult to predict at this time.

This transcriptome reveals a wealth of information about how the metabolic capacity in *Euglena* has evolved. This complex evolutionary history, combined with horizontal gene transfer (Henze et al., 1995), gives *Euglena* a huge biosynthetic capability, obtained from diverse sources, and has allowed the evolution of a complex genome which shows features of higher eukaryote control mechanisms. There is also evidence for evolution of novel enzymes, giving *Euglena* a unique metabolic competency. These observations are based on the 14,389 BLASTP hits, providing prospective functional annotation, leaving a further 17,739 non-redundant predicted proteins that show no significant homology to any known protein and for which we currently have no clue to their function.

Genome evolution

The kingdom-level distribution of the top BLASTP hits illustrates the huge diversity of sources of genetic material present in the *Euglena* genome, obtained from horizontal gene transfer, and highlights its complex genetic history (Fig. 1A).

The Euglenoids are related to the pathogenic protozoa Trypanosomes and Leishmania (Fig. 1B) and are extremely difficult to classify, even using modern molecular techniques (Linton et al., 2010). Since the split of *Euglena* from other members of the Euglenozoa over one billion years ago (Parfrey et al., 2011), there is evidence for a red algae endosymbiont, which transferred some genes to the eventual *Euglena* nuclear genome and has since been lost (Maruyama et al., 2011). Subsequently, there was endosymbiosis of a eukaryotic green alga (Martin et al., 1992), with transfer of many genes to the nucleus, including those for the maintenance of the chloroplast. Thus, the genetic material in *Euglena* is derived from: the ancestral protozoa; the mitochondrion, related to alpha-proteobacteria; the red algal endosymbiont; the primary photosynthetic host; and the primary photosynthetic endosymbiont, related to cyanobacteria, obtained from the green alga (Fig. 2).

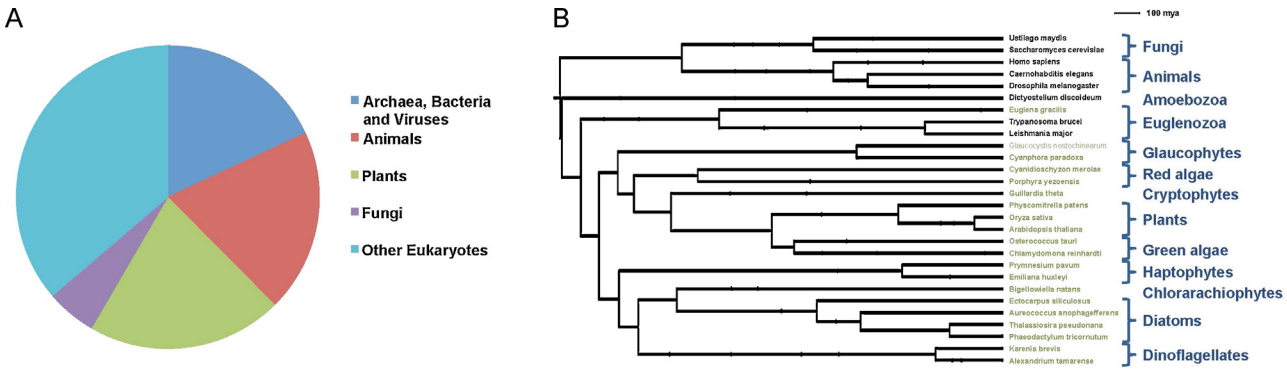

Figure 1 Genetic diversity in *Euglena*. (A). Kingdom-level taxonomic distribution of top hits of BLAST matches (*E*-values $<1 \times 10^{-10}$) of *Euglena gracilis* unique sequences. (B) Phylogeny of *E. gracilis* in relation to sequenced algae and model organisms (Letunic and Bork, 2011; Parfrey et al., 2011). Organisms in green are photosynthetic.

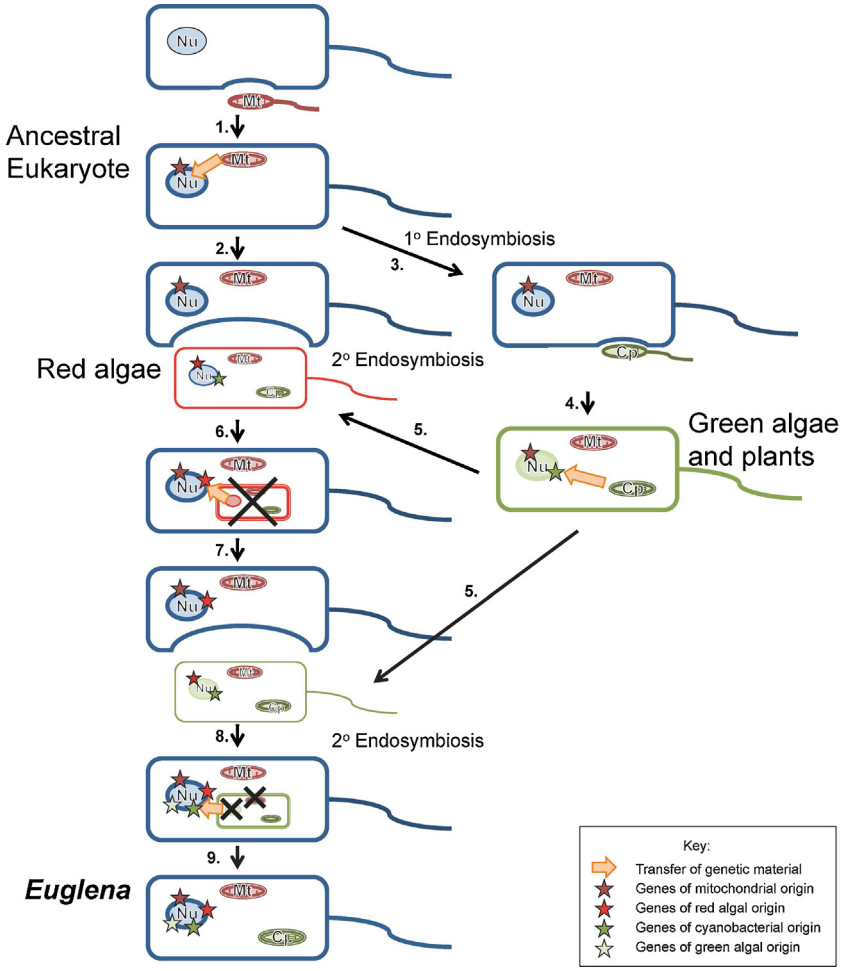

Figure 2 Sources of the *Euglena* genome. 1. The ancestor of all eukaryotic cells formed an endosymbiotic relationship with a prokaryote to form the mitochondrion (Mt). 2. Most of the genetic material was transferred to the nucleus (Nu) whilst some remains in the mitochondria. 3. The ancestor of all plant cells subsequently took up a cyanobacterial cell to form the chloroplast (Cp). 4. Most of the genetic material was transferred to the nucleus. 5. The ancestral plant cells diversified to form green algae and plants, golden algae and red algae (Moreira et al., 2000). 6. A red algal cell was taken up by the ancestor of Euglenids and some of the DNA transferred to the nucleus (Maruyama et al., 2011). 7. The red algae was then lost and the ancestor of photosynthetic *Euglena* formed an endosymbiotic relationship with a green algae (Gockel and Hachtel, 2000), which has been subsequently lost in several independent Euglenid lineages. 8. Nuclear and chloroplast DNA were subsequently transferred from the green algae to the nucleus of the Euglenid ancestor. 9. The nucleus and mitochondia were lost from the plant cell to leave the final chloroplast.

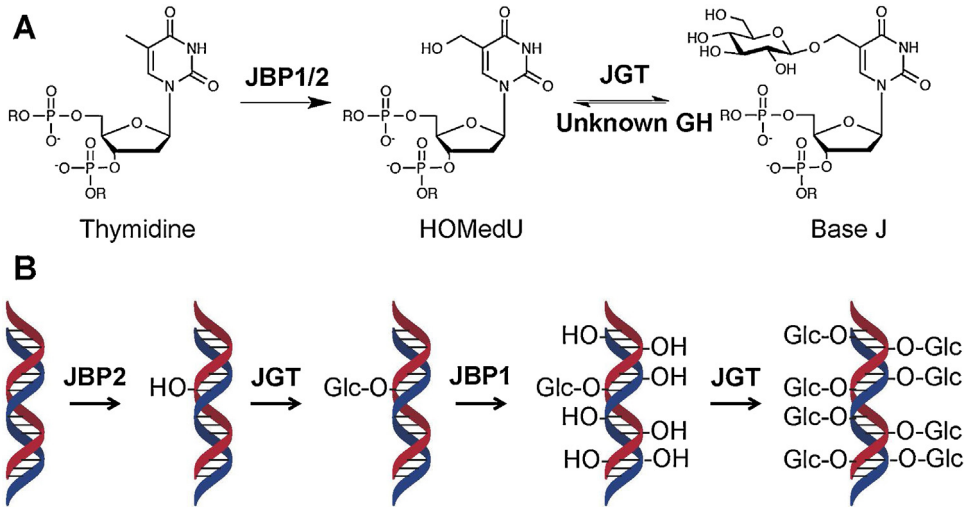

Figure 3 Base J metabolism. (A) To form Base J a thymidine is hydroxylated by JBPs, to form hydroxymethyl-deoxy-uracil (HOMedU), and then glucosylated by JGT, a novel glucosyltransferase. The glucose is removed by an as yet unidentified glucosidase, leaving HOMedU. R_1 = 3′ DNA strand, R_2 = 5′ DNA strand. (B) JBP2 is thought to initiate *de novo* hydroxylation of thymidine. The product is then glucosylated by JGT to form novel Base J residues, which are recognised by JBP1, leading to hydroxylation of nearby thymidine residues. Further glucosylation by JGT subsequently leads to a local amplification of the Base J silencing signal (Borst and Sabatini, 2008).

Despite being colloquially referred to as a green alga, the core nuclear genome of *Euglena* is more closely related to that of Trypanosomes than to that of other eukaryotic algae.

In addition to the nuclear genome, some genetic material is retained in the chloroplast (Hallick et al., 1993) and the mitochondrion, which has an unusually fragmented gene organisation (Spencer and Gray, 2011). The Euglenoid chloroplast has had substantial genetic rearrangement during its evolution (Hrdá et al., 2012). This has included the transfer of several important genes from the chloroplast genome to the nucleus, from where they can be identified in the *Euglena* transcriptome, including the protease clpP (lm.95241 and lm.15675), the membrane protein cemA (lm.59206) and photosystem 1 component ycf3 (lm.46611).

Evolution of controls

To control expression and activity of proteins eukaryotes use sophisticated mechanisms, with more intricate systems amongst the higher organisms. *Euglena* has many of the classical mechanisms, as well as some more unusual and more complex than have been found elsewhere.

Base J

Base J is a modified DNA base, formed by hydroxylation and glucosylation of thymidine (Fig. 3A). It is uniquely found amongst the Euglenozoa, including the important human pathogens Trypanosomes and Leishmania (van Leeuwen et al., 1998). It prevents RNA polymerases passing along the DNA strand, silencing the modified region of the genome, and is of key importance to the control of surface coat protein expression in Trypanosomes (Borst and Sabatini, 2008). In Leishmania, some of the Base J is not located around the telomeres, as is the case in Trypanosomes (van Luenen et al., 2012), and it has been shown to prevent transcriptional read-through, regulating transcription termination (Reynolds et al., 2014). In contrast, Base J in *Euglena* is found throughout the genome where it makes up approximately 0.2% (1 in 500) of the bases (Dooijes et al., 2000).

The two proteins involved in the initial hydroxylation of thymidine are well studied in Trypanosomes and Leishmania (DiPaolo et al., 2005). There are matching homologues encoded in the *Euglena* transcriptome, though the JBP1 homologue (dm.72228), essential in Leishmania, (Genest et al., 2005) is not found in the transcriptome of the light grown cells. Unlike JBP1, JBP2 (dm.12798 in *Euglena*) does not actually bind Base J, but instead appears to initiate *de novo* Base J biosynthesis, whilst JBP1 amplifies this signal (Fig. 3B) (Cliffe et al., 2009).

Recently, two separate groups identified the Base J glucosyltransferase as a single copy in the Trypanosome genomes (Bullard et al., 2014; Sekar et al., 2014). This enzyme was noted as being distantly related to GT-A type glycosyltransferases and to have variable loops between conserved regions in other kinetoplastid species. In the *Euglena* transcriptome there is one homologue (dm.53028) of the Trypanosome transferase, with an alternative splice variant containing a short N-terminal region of unclear function (dm.53027). The enzyme(s) for removal of the glucose unit have not, to date, been identified in any organism. After removal of the glucose the HOMedU could be either dehydroxylated or the base could be excised and replaced (Borst and Sabatini, 2008).

Gene silencing

RNA-mediated gene silencing is a ubiquitous control mechanism found in bacteria (Marraffini and Sontheimer, 2010), plants (Baulcombe, 2004) and animals (Berezikov and Plasterk, 2005). Three main components make up the machinery in eukaryotes: Dicer like (DCL), which cleaves

Table 1 Components of the RNA silencing machinery. Transcripts for genes involved in gene silencing pathways were identified in the *Euglena* transcriptome (Brodersen and Voinnet, 2006). tasi: transacting siRNA. S-PTGS: sense posttranscriptional gene silencing. IR-PTGS: inverted repeat posttranscriptional gene silencing.

Protein name	Activity	Pathway	No of homologues in *Euglena*
AGO	RNA slicer	miRNA, S-PTGS, tasi-RNA, chromatin	4
DCL	miRNA synthesis	All	4
RDR	RNA-dependant RNA-polymerase	S-PTGS, transitivity, tasi-RNA, nat-siRNA, chromatin	0
CMT	Cytosine DNA methyltransferase	Chromatin	2
K9 MeT	Histone methyl transferase	Chromatin	5
NRPD	DNA-dependant RNA-polymerase	Chromatin, nat-siRNA	8
HDA6	Histone deacetylase	Chromatin	8
SDE3	Helicase	S-PTGS, transitivity	8
HST	Exportin	miRNA	1
HYL1	dsRNA binding	miRNA	0
WEX	Exonuclease	S-PTGS	0
SGS3	Unknown	S-PTGS, transitivity, tasiRNA, nat-siRNA	0
HEN1	sRNA-methyl transferase	All	0

double stranded RNA (Liu et al., 2009); Argonaute (AGO), which targets inactivation of sequences complementary to small RNAs (21-24nt) as part of the RNA-induced silencing complex (RISC) (Hutvagner and Simard, 2008); and RNA-dependent RNA polymerase (RDRP), which amplifies the silencing (Baulcombe, 2007). Whilst some sequenced algae do not have any of the necessary components, most retain some capability for gene silencing (Cerutti et al., 2011). *Euglena* is known to possess some capacity for RNA-silencing: successful RNAi knockdown experiments of a gene involved in vitamin C biosynthesis (Ishikawa et al., 2008) and a photoreceptor protein (Ntefidou et al., 2003) have been performed in this organism. *Arabidopsis* encodes four copies of DCL, which are differentially expressed under stress conditions (Liu et al., 2009). The *Euglena* transcriptome also encodes four DCL proteins, which appear to have diverged from a single copy rather than being acquired by repeated horizontal gene transfer or endosymbiosis. Argonautes are split between the piwi subfamily and the AGO subfamily, which have undergone duplication, expansion and loss in different lineages (Hutvagner and Simard, 2008). Humans have four genes for each subfamily, whilst *Arabidopsis* has ten members of the AGO subfamily and no piwi-like genes, and Leishmania has lost both of these subfamilies entirely. There are no members of the piwi subfamily encoded in the *Euglena* transcriptome, but there are four AGO family proteins, with three being closely related and one more divergent, possibly suggesting two separate gene acquisitions.

There are also many of the other components of the gene silencing machinery in the *Euglena* transcriptome, including helicases, histone methylases, histone deacetylases, and cytosine methylases (Table 1) and these are variably expressed in the light-grown and dark-grown cells. There is no RNA-dependent RNA polymerase (RDRP) present, which suggests that *Euglena* is incapable of producing trans-acting RNAs or utilising the aberrant transcript pathway (Brodersen and Voinnet, 2006). It is possible that another enzyme is capable of acting to amplify the signal, as is suggested for mammals and insects, which also lack the RDRP (Baulcombe, 2007). In *Arabidopsis* this amplification is involved in development, where the original double-stranded RNA might be diluted during growth (Fahlgren et al., 2006), which is conceivably not required in a single celled organism such as *Euglena*.

O-GlcNAc

As a complement to protein phosphorylation, a key regulator of protein activity is the addition and removal of serine and threonine linked *N*-acetylglucosamine (O-GlcNAc) (Love and Hanover, 2005). This modification impacts on cellular signalling and nutrient response, including cross-talk with protein phosphorylation, which competes for the same sites on the protein (Fig. 4) (Zeidan and Hart, 2010). There are three putative GT41 O-GlcNAc transferases in the *Euglena* transcriptome, (lm.52466, dm.35031 and lm.92993), the latter of which is only present in the light grown cells. In humans there is only one transferase gene (Vocadlo, 2012), whilst plants have two distinct enzymes, SEC, similar to the animal enzyme, and SPY (Olszewski et al., 2010). There are, however, no homologues of the human O-GlcNAcase encoded in the *Euglena* transcriptome, or in plants, which would reverse the O-GlcNAc addition, suggesting that a non-canonical hydrolase may carry out this reaction. Hence it is evident that not only does *Euglena* carry out this 'higher-eukaryote' protein glycosylation, it employs a more complex system than 'higher' multi-cellular organisms.

Evolution of protein architecture

Euglena has enzymes for the biosynthesis of many diverse compounds, including amino acids, vitamins, complex

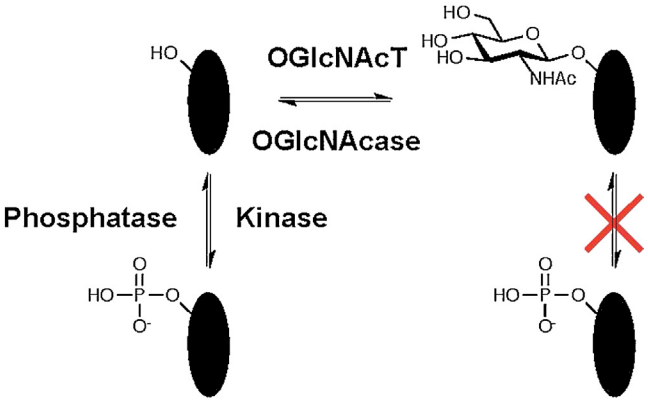

Figure 4 Protein modification by O-GlcNAc. OGlcNAcT transfers GlcNAc onto serine and threonine residues and OGlcNAcase removes it (Vocadlo, 2012). This modification is orthogonal to kinase-mediated phosphorylation.

carbohydrates and polyunsaturated fatty acids (O'Neill et al., 2015). These capabilities have been obtained from many diverse sources through evolution. Aside from mutation-based evolution, such as the massive expansion and diversification of carbohydrate-active enzyme families (Cantarel et al., 2009), *Euglena* appears to have made extensive use of gene fusions to produce novel domain arrangements, a few examples of which are discussed below.

Alternative splicing

Many of the transcripts obtained from sequencing represent alternative splicing variants, information that would not be available from genome sequencing. For example, transcripts lm.75841 and lm.75842 share an identical N-terminus, coding for a glycosyltransferase, but the former has a C-terminal extension not present in the latter, which encodes a peroxisomal protein (Fig. 5A). The ratio of the expression of the long to short isoforms was estimated to be approximately 11:1 in the light sample, but in the dark no short sequence variant is detectable. This suggests that in the light the enzyme activity is required in both subcellular locations, but in the dark it is not required in the cytosol. Hence *Euglena* appears to make use of alternative splicing to control subcellular targeting of a single gene product, as has been seen for many enzymes (Danpure, 1995), including glycolytic enzymes in fungi (Freitag et al., 2012) and amino acid metabolic enzymes in plants (Gebhardt et al., 1998).

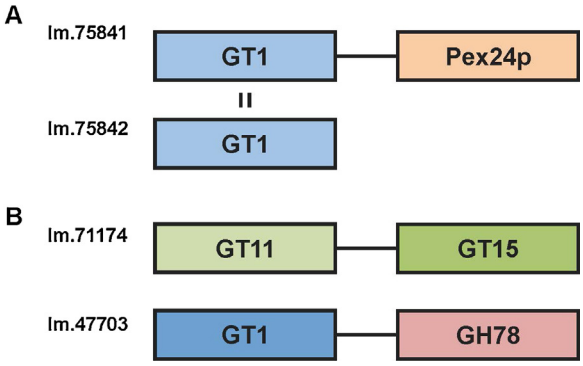

Figure 5 (A) Examples of splice variants in *Euglena* transcripts. lm.75841 and lm.75842 are identical for the first 354 amino acids, comprising a GT1 domain, but lm.75841 has a further 421 amino acids, including a domain related to Pex24p, an integral peroxisomal membrane protein. (B) Domain architecture of didomain CAZys encoded in the *Euglena* transcriptome lm.71174 encodes an N-terminal GT11 and a C-terminal GT15. dm.47703 encodes an N-terminal GT1 and a C-terminal GH78.

Di-domain carbohydrate-active enzymes

Whilst fusion of carbohydrate binding modules (CBMs) to carbohydrate-active enzymes is relatively common in nature, and contiguity of multiple glycoside hydrolase domains in a single protein is well known, it is much rarer to find glycosyltransferases as part of a protein containing other domains. Examples include the sea-squirt *Oikopleura dioica* (Medie et al., 2012), which encodes a cellulose synthase (GT2) and a β-glucan hydrolase (GH6), and the previously discussed O-GlcNAc transferases, which are found in most eukaryotes and some bacteria, and encode tetratricopeptide repeats in addition to the GT41 GlcNAc-transferase domain (Lubas et al., 1997).

Two transcripts were identified in the *Euglena* transcriptome that encode proteins with two carbohydrate-active enzyme domains (Fig. 5B). The first protein (lm.71174) has a putative GT11 fucosyltransferase domain and a putative GT15 mannosyltransferase domain. The active site of the former does not contain the second arginine in the HxRRxD motif (Takahashi et al., 2000), whilst the latter contains the nucleophile and a zwitterionic ion-binding motif (Lobsanov et al., 2004). It is possible that this enzyme maybe act to transfer both fucose and mannose to the same *N*-glycan core. Alternatively, this enzyme may transfer mannose on to a fucosylated glycan, to which the GT11 domain, acting as a carbohydrate-binding module, directs it.

A second di-domain protein (dm.47703) is composed of a GT1 sugar transferase, most closely related to bacterial sterol β-glucuronic acid transferases, linked to a C-terminal GH78 α-rhamnosidase domain. Both domains appear to have an intact active site, suggesting that both activities are viable (Cui et al., 2007; Mulichak et al., 2004). This di-domain protein might conceivably be involved in cleaving rhamnose from a small molecule and adding a glucuronic acid moiety, a sugar addition that is known to facilitate subcellular relocalisation and xenobiotic detoxification (Tukey and Strassburg, 2000).

Tryptophan biosynthesis

The biosynthesis of tryptophan from chorismate, *via* anthranilate, is typically carried out by five sequential enzymatic reactions (Fig. 6) (Crawford, 1975). The first reaction is catalysed by a two-component anthranilate synthase, though these proteins are often fused. In *Euglena*, there are two transcripts encoding both components, one of which contains an additional N-terminal aminotransferase, which has not previously been noted. The next four reactions in the biosynthetic pathway are normally located on separate proteins, though there is sometimes

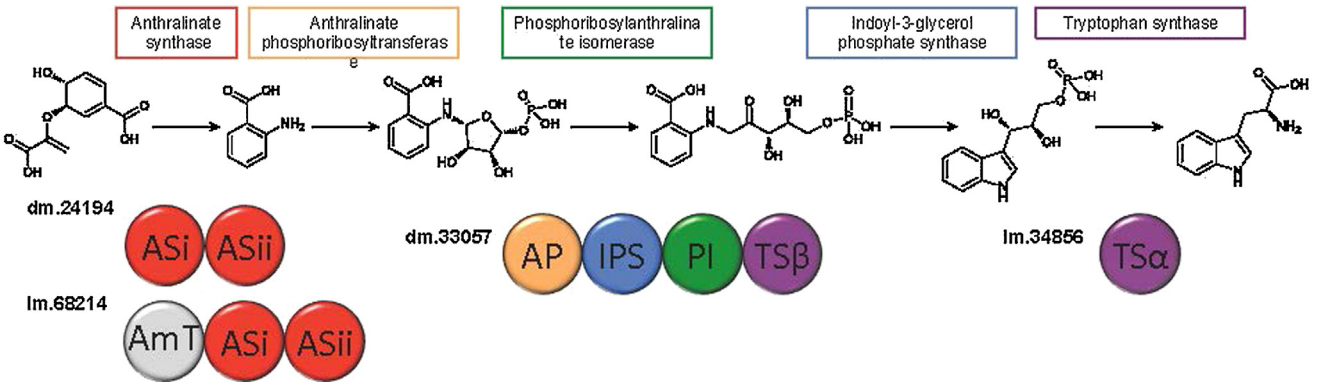

Figure 6 Tryptophan biosynthesis. The domains are typically encoded as separate proteins but there may sometimes been two domain fusions. In *Euglena* there are two transcripts encoding both components of the anthranilate synthase, one of which also has an extra aminotransferase domain of unknown function. A novel four-domain enzyme carries out the biosynthesis of tryptophan from anthralinate, with a separately encoded tryptophan synthase α-domain.

one gene fusion (Cohn et al., 1979). Uniquely in *Euglena*, a single transcript encodes all four enzymes, which has not previously been observed (Schwarz et al., 1997). This unusual fusion construct has three domains related to fungal enzymes and a bacterial isomerase; the domain sequence is not in biosynthetic order. The final synthase, required for indole formation and condensation to serine, is composed of a hetero-tetramer in bacteria (α_2,β_2) (Miles, 2006) and a homo-dimer in fungi ($(\alpha\beta)_2$) (Matchett and DeMoss, 1975). The tetra-functional enzyme from the *Euglena* transcriptome only contains the β-subunit, whilst the α-chain is encoded on a separate transcript. This suggests that tryptophan biosynthesis is catalysed by a hetero-tetramer composed of two copies of the multi-function protein and two copies of the tryptophan synthase α-chain. This once again highlights the unique capabilities that *Euglena* displays.

Natural product synthases

No polyketides or non-ribosomal peptides have been confirmed in *Euglena* to date. However, there are transcripts apparent for the complex multi-domain secondary metabolite synthases needed to make such compounds, as is evident for an increasing array of algae now that genome/transcriptome sequence data is becoming available (Sasso et al., 2012).

Polyketides comprise a huge range of compounds, formed by repeated condensation of acetate units, followed by variable reduction and further elaboration. Broadly speaking, polyketide synthases (PKSs) can be large multi-domain proteins (type I) or composed of discrete proteins with individual functions (type II), although there are other architectures possible (Shen et al., 2007). Fourteen potential PKSs were identified in *Euglena* as having the key ketosynthase domain and attempts to predict the structures of the compounds synthesised by these, using SBSPKS (Anand et al., 2010) and the PKS/NRPS Analysis Web-site (Bachmann and Ravel, 2009), were not successful. Analysis of the domain sequences of these enzymes using DELTA-BLAST allows some putative predictions to be made. For example, the largest polyketide synthase encoded in the *Euglena* transcriptome (lm.8157) contains, in addition to a fully reducing PKS module, two enoyl hydratases and an HMGCoA synthase (Fig. 7A). These proteins have been characterised in bacterial gene clusters as single domain proteins, such as PksG, H and I in bacillaene biosynthesis, as adding a β-methyl branch to polyketides (Butcher et al., 2007). Whilst the association of other components of the complete synthase cannot be predicted, this domain architecture suggests the formation of methyl branched alkane, which could be part of a polyketide, or alternatively may be included in a fatty acid structure.

Non-ribosomal peptide synthetases (NRPSs) are multi-domain proteins that join amino acids together to form small peptides with a diversity of function. Five proteins encoded in the *Euglena* transcriptome contain both A-domains, for the activation of the amino acids, and C-domains, for formation of the peptide bonds, with the PCP acting as a carrier for the growing peptide. For example, lm.32232 contains a complete module in the characteristic C-A-PCP domain order, followed by an additional module lacking the PCP and an extra N-terminal C-domain with no associated A

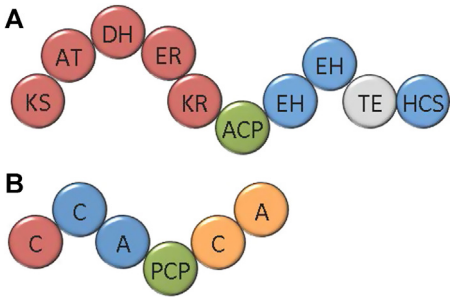

Figure 7 Multidomain natural product megasynthases in *Euglena*. (A) lm.8157 is composed of 5 modules of a fully reducing polyketide synthase (in red) followed by three modules (in blue) that add a β-methyl branch to polyketides and a thioesterase domain (in grey) for release of the product from the acyl carrier protein (green). (B) lm 3223 is a non-ribosomal peptide synthase with two adenylation domains for activation of specific amino acids and three condensation domains for formation of peptide bonds.

domain or PCP, which is occasionally seen in microbial NRPSs. The A-domains specify the amino acid but the prediction programmes SBSPKS (Anand et al., 2010) and NRPS/PKS Analysis Web-site (Bachmann and Ravel, 2009), are incapable of dealing with these sequences, probably because of the evolutionary distance from the bacterial and fungal species with which these pieces of software were designed to deal.

Conclusions

The transcriptome of *Euglena* reveals a huge diversity of proteins, many of which have no known equivalents in other species. Although only a unicellular organism, the number of genes and the control mechanisms evident in *Euglena* are as sophisticated as those typically found in higher eukaryotes, if not more so. A diverse array of metabolic enzymes has been acquired through the complex evolutionary history of *Euglena* over the last 1.6 billion years, since their divergence from plants and other algae. A number of *Euglena* genes encode fusions of several domains encoded on one polypeptide, acquired either through horizontal gene transfer from microorganisms or generated as novel fusions within *Euglena*. This suggests that there is some selective pressure that favours placing several domains from single pathways on one protein, rather than relying on diffusion or assembling multi component complexes non-covalently. Perhaps the large cytoplasmic volume of *Euglena* cells, with a highly flexible cell shape and no vacuole, renders the diffusion of metabolites or proteins insufficient for efficient metabolism.

The novel and complex evolution of *Euglena* provides a wealth of unexpected information accessed from its transcriptome. There are many unusual features in genome and transcriptome dynamics, including mechanisms for metabolic control that are more complex than in 'higher' multi-cellular organisms. In addition, *Euglena* offers major opportunities for metabolic engineering and for the production of added-value biomolecules (vitamins, natural products, essential amino acids), and provides inspiration for the transfer of its diverse metabolic capabilities into other host systems.

Conflict of interest

The authors declare that there is no conflict of interest.

Acknowledgments

These studies were supported by the UK BBSRC Institute Strategic Programme Grant on Understanding and Exploiting Metabolism (MET) [BB/J004561/1] and the John Innes Foundation.

References

Anand, S., Prasad, M.V.R., Yadav, G., Kumar, N., Shehara, J., Ansari, M.Z., Mohanty, D., 2010. SBSPKS: structure based sequence analysis of polyketide synthases. Nucleic Acids Res. 38 (Suppl. 2), W487—W496, http://dx.doi.org/10.1093/nar/gkq340.

Bachmann, B.O., Ravel, J., 2009. Methods for in silico prediction of microbial polyketide and nonribosomal peptide biosynthetic pathways from DNA sequence data. In: Hopwood, D.A. (Ed.), Complex enzymes in microbial natural product biosynthesis, Part A: Overview articles and peptides. Methods in Enzymology, vol. 458, pp. 181—217, http://dx.doi.org/10.1016/S0076-6879(09)04808-3.

Baulcombe, D., 2004. RNA silencing in plants. Nature 431 (7006), 356—363, http://dx.doi.org/10.1038/nature02874.

Baulcombe, D.C., 2007. Amplified silencing. Science 315 (5809), 199—200, http://dx.doi.org/10.1126/science.1138030.

Berezikov, E., Plasterk, R.H.A., 2005. Camels and zebrafish, viruses and cancer: a microRNA update. Hum. Mol. Genet. 14 (Suppl. 2), R183—R190, http://dx.doi.org/10.1093/hmg/ddi271.

Borst, P., Sabatini, R., 2008. Base J: discovery, biosynthesis, and possible functions. Annu. Rev. Microbiol. 621, 235—251, http://dx.doi.org/10.1146/annurev.micro.62.081307.162750.

Brodersen, P., Voinnet, O., 2006. The diversity of RNA silencing pathways in plants. Trends Genet. 225, 268—280, http://dx.doi.org/10.1016/j.tig.2006.03.003.

Bullard, W., da Rosa-Spiegler, J.L., Liu, S., Wang, Y., Sabatini, R., 2014. Identification of the glucosyltransferase that converts hydroxymethyluracil to Base J in the trypanosomatid genome. J. Biol. Chem. 289, 20273—20282, http://dx.doi.org/10.1074/jbc.m114.579821.

Butcher, R.A., Schroeder, F.C., Fischbach, M.A., Straight, P.D., Kolter, R., Walsh, C.T., Clardy, J., 2007. The identification of bacillaene, the product of the PksX megacomplex in *Bacillus subtilis*. Proc. Natl. Acad. Sci. U. S. A. 1045, 1506—1509, http://dx.doi.org/10.1073/pnas.0610503104.

Cantarel, B.L., Coutinho, P.M., Rancurel, C., Bernard, T., Lombard, V., Henrissat, B., 2009. The Carbohydrate-Active EnZymes database (CAZy): an expert resource for glycogenomics. Nucleic Acids Res. 37, D233—D238, http://dx.doi.org/10.1093/nar/gkn663.

Cerutti, H., Ma, X., Msanne, J., Repas, T., 2011. RNA-mediated silencing in algae: biological roles and tools for analysis of gene function. Eukaryot. Cell 109, 1164—1172, http://dx.doi.org/10.1128/ec.05106-11.

Cliffe, L.J., Kieft, R., Southern, T., Birkeland, S.R., Marshall, M., Sweeney, K., Sabatini, R., 2009. JBP1 and JBP2 are two distinct thymidine hydroxylases involved in J biosynthesis in genomic DNA of African trypanosomes. Nucleic Acids Res. 375, 1452—1462, http://dx.doi.org/10.1093/nar/gkq146.

Cohn, W., Kirschner, K., Paul, C., 1979. N-(5-Phosphoribosyl) anthranilate isomerase-indoleglycerol-phosphate synthase 2. Fast-reaction studies show that a fluorescent substrate analog binds independently to two different sites. Biochemistry 1826, 5953—5959, http://dx.doi.org/10.1021/bi00593a030.

Crawford, I.P., 1975. Gene rearrangements in evolution of tryptophan synthetic pathway. Bacteriol. Rev. 392, 87—120.

Cui, Z., Maruyama, Y., Mikami, B., Hashimoto, W., Murata, K., 2007. Crystal structure of glycoside hydrolase family 78 α-L-rhamnosidase from *Bacillus* sp. GL1. J. Mol. Biol. 3742, 384—398, http://dx.doi.org/10.1016/j.jmb.2007.09.003.

Danpure, C.J., 1995. How can the products of a single gene be localized to more than one intracellular compartment? Trends Cell Biol. 56, 230—238, http://dx.doi.org/10.1016/s0962-8924(00)89016-9.

DiPaolo, C., Kieft, R., Cross, M., Sabatini, R., 2005. Regulation of trypanosome DNA glycosylation by a SWI2/SNF2-like protein. Mol. Cell 173, 441—451, http://dx.doi.org/10.1016/j.molcel.2004.12.022.

Dooijes, D., Chaves, I., Kieft, R., Dirks-Mulder, A., Martin, W., Borst, P., 2000. Base J originally found in Kinetoplastida is also a minor constituent of nuclear DNA of *Euglena gracilis*. Nucleic Acids Res. 2816, 3017—3021, http://dx.doi.org/10.1093/nar/28.16.3017.

Fahlgren, N., Montgomery, T.A., Howell, M.D., Allen, E., Dvorak, S.K., Alexander, A.L., Carrington, J.C., 2006. Regulation of AUXIN RESPONSE FACTOR3 by TAS3 ta-siRNA affects developmental timing and patterning in *Arabidopsis*. Curr. Biol. 169, 939–944, http://dx.doi.org/10.1016/j.cub.2006.03.065.

Freitag, J., Ast, J., Bolker, M., 2012. Cryptic peroxisomal targeting via alternative splicing and stop codon read-through in fungi. Nature 485 (7399), 522–525, http://dx.doi.org/10.1038/nature11051.

Gebhardt, J.S., Wadsworth, G.J., Matthews, B.F., 1998. Characterization of a single soybean cDNA encoding cytosolic and glyoxysomal isozymes of aspartate aminotransferase. Plant Mol. Biol. 371, 99–108.

Genest, P.-A., ter Riet, B., Dumas, C., Papadopoulou, B., van Luenen, H.G.A.M., Borst, P., 2005. Formation of linear inverted repeat amplicons following targeting of an essential gene in *Leishmania*. Nucleic Acids Res. 335, 1699–1709, http://dx.doi.org/10.1093/nar/gki304.

Gockel, G., Hachtel, W., 2000. Complete gene map of the plastid genome of the nonphotosynthetic euglenoid flagellate *Astasia longa*. Protist 1514, 347–351, http://dx.doi.org/10.1007/bf00020882.

Hallick, R.B., Hong, L., Drager, R.G., Favreau, M.R., Monfort, A., Orsat, B., Spielmann, A., Stutz, E., 1993. Complete sequence of *Euglena gracilis* chloroplast DNA. Nucleic Acids Res. 2115, 3537–3544, http://dx.doi.org/10.1093/nar/21.15.3537.

Henze, K., Badr, A., Wettern, M., Cerff, R., Martin, W., 1995. A nuclear gene of eubacterial origin in *Euglena gracilis* reflects cryptic endosymbioses during protist evolution. Proc. Natl. Acad. Sci. U. S. A. 9220, 9122–9126, http://dx.doi.org/10.1073/pnas.92.20.9122.

Hutvagner, G., Simard, M.J., 2008. Argonaute proteins: key players in RNA silencing. Nat. Rev. Mol. Cell Biol. 91, 22–32, http://dx.doi.org/10.1038/nrm2321.

Hrdá, Š., Fousek, J., Szabová, J., Hampl, V.V., Vlček, Č., 2012. The plastid genome of *Eutreptiella* provides a window into the process of secondary endosymbiosis of plastid in Euglenids. PLoS ONE 73, e33746, http://dx.doi.org/10.1371/journal.pone.0033746.

Inui, H., Miyatake, K., Nakano, Y., Kitaoka, S., 1982. Wax ester fermentation in *Euglena gracilis*. FEBS Lett. 1501, 89–93, http://dx.doi.org/10.1016/0014-5793(82)81310-0.

Ishikawa, T., Nishikawa, H., Gao, Y., Sawa, Y., Shibata, H., Yabuta, Y., Maruta, T., Shigeoka, S., 2008. The pathway via D-galacturonate/L-galactonate is significant for ascorbate biosynthesis in *Euglena gracilis*. J. Biol. Chem. 283 (45), 31133–31141, http://dx.doi.org/10.1074/jbc.m803930200.

Korn, E.D., 1964. The fatty acids of *Euglena gracilis*. J. Lipid Res. 53, 352–362.

Letunic, I., Bork, P., 2011. Interactive Tree of Life v2: online annotation and display of phylogenetic trees made easy. Nucleic Acids Res. 39 (Suppl. 2), W475–W478, http://dx.doi.org/10.1093/nar/gkr201.

Linton, E.W., Karnkowska-Ishikawa, A., Kim, J.I., Shin, W., Bennett, M.S., Kwiatowski, J., Zakrys, B., Triemer, R.E., 2010. Reconstructing Euglenoid evolutionary relationships using three genes: nuclear SSU and LSU, and chloroplast SSU rDNA sequences and the description of euglenaria gen. nov (Euglenophyta). Protist 1614, 603–619, http://dx.doi.org/10.1016/j.protis.2010.02.002.

Liu, Q., Feng, Y., Zhu, Z., 2009. Dicer-like (DCL) proteins in plants. Funct. Integr. Genomics 93, 277–286, http://dx.doi.org/10.1007/s10142-009-0111-5.

Lobsanov, Y.D., Romero, P.A., Sleno, B., Yu, B., Yip, P., Herscovics, A., Howell, P.L., 2004. Structure of Kre2p/Mnt1p: a yeast α-1,2-mannosyltransferase involved in mannoprotein biosynthesis. J. Biol. Chem. 27917, 17921–17931.

Love, D.C., Hanover, J.A., 2005. The Hexosamine signaling pathway: deciphering the ''O-GlcNAc Code''. Sci. STKE 2005 (312), re13, http://dx.doi.org/10.1126/stke.3122005re13.

Lubas, W.A., Frank, D.W., Krause, M., Hanover, J.A., 1997. O-linked GlcNAc transferase is a conserved nucleocytoplasmic protein containing tetratricopeptide repeats. J. Biol. Chem. 27214, 9316–9324.

Marraffini, L.A., Sontheimer, E.J., 2010. CRISPR interference: RNA-directed adaptive immunity in bacteria and archaea. Nat. Rev. Genet. 113, 181–190, http://dx.doi.org/10.1038/nrg2749.

Martin, W., Somerville, C.C., Loiseaux-de Goer, S., 1992. Molecular phylogenies of plastid origins and algal evolution. J. Mol. Evol. 355, 385–404, http://dx.doi.org/10.1007/bf00171817.

Maruyama, S., Suzaki, T., Weber, A.P.M., Archibald, J.M., Nozaki, H., 2011. Eukaryote-to-eukaryote gene transfer gives rise to genome mosaicism in Euglenids. BMC Evol. Biol. 11, 105, http://dx.doi.org/10.1186/1471-2148-11-105.

Matchett, W.H., DeMoss, J.A., 1975. The subunit structure of tryptophan synthase from *Neurospora crassa*. J. Biol. Chem. 2508, 2941–2946.

Medie, F.M., Davies, G.J., Drancourt, M., Henrissat, B., 2012. Genome analyses highlight the different biological roles of cellulases. Nat. Rev. Microbiol. 103, 227–234, http://dx.doi.org/10.1038/nrmicro2729.

Miles, E.W., 2006. Structural basis for catalysis by tryptophan synthase. In: Meister, A. (Ed.), Adv. Enzymology and Related Areas of Molecular Biology, vol. 64. John Wiley & Sons, Inc., Hoboken, NJ, USA, pp. 93–172, http://dx.doi.org/10.1002/9780470123102.ch3.

Moreira, D., Le Guyader, H., Philippe, H., 2000. The origin of red algae and the evolution of chloroplasts. Nature 405 (6782), 69–72, http://dx.doi.org/10.1038/35011054.

Mulichak, A.M., Lu, W., Losey, H.C., Walsh, C.T., Garavito, R.M., 2004. Crystal structure of vancosaminyltransferase GtfD from the vancomycin biosynthetic pathway: interactions with acceptor and nucleotide ligands. Biochemistry 4318, 5170–5180, http://dx.doi.org/10.1021/bi036130c.

Ntefidou, M., Iseki, M., Watanabe, M., Lebert, M., Häder, D.-P., 2003. Photoactivated adenylyl cyclase controls phototaxis in the flagellate *Euglena gracilis*. Plant Physiol. 133, 1517–1521, http://dx.doi.org/10.1104/pp.103.034223.

Olszewski, N.E., West, C.M., Sassi, S.O., Hartweck, L.M., 2010. O-GlcNAc protein modification in plants: evolution and function. Biochim. Biophys. Acta Gen. Subj. 1800 (2), 49–56, http://dx.doi.org/10.1016/j.bbagen.2009.11.016.

O'Neill, E.C., Trick, M., Hill, L., Rejzek, M., Dusi, R.G., Hamilton, C.J., Zimba, V., Henrissat, B., Field, R.A., 2015. The transcriptome of Euglena gracilis reveals unexpected metabolic capabilities for carbohydrate and natural product biochemistry. Mol. Biosyst. 11, 2808–2820, Available at: http://jicbio.nbi.ac.uk/euglena/.

Parfrey, L.W., Lahr, D.J.G., Knoll, A.H., Katz, L.A., 2011. Estimating the timing of early eukaryotic diversification with multigene molecular clocks. Proc. Natl. Acad. Sci. U. S. A. 108 (33), 13624–13629, http://dx.doi.org/10.1073/pnas.1110633108.

Reynolds, D., Cliffe, L., Foerstner, K.U., Hon, C.-C., Siegel, T.N., Sabatini, R., 2014. Regulation of transcription termination by glucosylated hydroxymethyluracil, base J, in *Leishmania major* and *Trypanosoma brucei*. Nucleic Acids Res. 4215, 9717–9729, http://dx.doi.org/10.1093/nar/gku714.

Rismani-Yazdi, H., Haznedaroglu, B.Z., Bibby, K., Peccia, J., 2011. Transcriptome sequencing and annotation of the microalgae *Dunaliella tertiolecta*: pathway description and gene discovery for production of next-generation biofuels. BMC Genomics 12, 148, http://dx.doi.org/10.1186/1471-2164-12-148.

Rodríguez-Zavala, J.S., Ortiz-Cruz, M.A., Mendoza-Hernández, G., Moreno-Sánchez, R., 2010. Increased synthesis of α-tocopherol, paramylon and tyrosine by *Euglena gracilis* under conditions of

high biomass production. J. Appl. Microbiol. 1096, 2160—2172, http://dx.doi.org/10.1111/j.1365-2672.2010.04848.x.

Sasso, S., Pohnert, G., Lohr, M., Mittag, M., Hertweck, C., 2012. Microalgae in the postgenomic era: a blooming reservoir for new natural products. FEMS Microbiol. Rev. 364, 761—785, http://dx.doi.org/10.1111/j.1574-6976.2011.00304.x.

Schwarz, T., Uthoff, K., Klinger, C., Meyer, H.E., Bartholmes, P., Kaufmann, M., 1997. Multifunctional tryptophan-synthesizing enzyme: the molecular weight of the *Euglena gracilis* protein is unexpectedly low. J. Biol. Chem. 27216, 10616—10623.

Sekar, A., Merritt, C., Baugh, L., Stuart, K., Myler, P.J., 2014. Tb927.10.6900 encodes the glucosyltransferase involved in synthesis of base J in *Trypanosoma brucei*. Mol. Biochem. Parasitol. 1961, 9—11, http://dx.doi.org/10.1016/j.molbiopara.2014.07.005.

Shen, B., Cheng, Y.-Q., Christenson Steven, D., Jiang, H., Ju, J., Kwon, H.-J., Lim, S.-K., Liu, W., Nonaka, K., Seo, J.-W., Smith Wyatt, C., Standage, S., Tang, G.-L., Van Lanen, S., Zhang, J., 2007. Polyketide Biosynthesis beyond the Type I, II, and III Polyketide Synthase Paradigms: A Progress Report, vol. 955. Polyketides, American Chemical Society, pp. 154—166, http://dx.doi.org/10.1021/bk-2007-0955.ch011.

Spencer, D., Gray, M., 2011. Ribosomal RNA genes in *Euglena gracilis* mitochondrial DNA: fragmented genes in a seemingly fragmented genome. Mol. Genet. Genomics 2851, 19—31, http://dx.doi.org/10.1007/s00438-010-0585-9.

Takahashi, T., Ikeda, Y., Tateishi, A., Yamaguchi, Y., Ishikawa, M., Taniguchi, N., 2000. A sequence motif involved in the donor substrate binding by α-1,6-fucosyltransferase: the role of the conserved arginine residues. Glycobiology 105, 503—510, http://dx.doi.org/10.1093/glycob/10.5.503.

Takeyama, H., Kanamaru, A., Yoshino, Y., Kakuta, H., Kawamura, Y., Matsunaga, T., 1997. Production of antioxidant vitamins β-carotene, vitamin C, and vitamin E, by two-step culture of *Euglena gracilis* Z. Biotechnol. Bioeng. 532, 185—190, http://dx.doi.org/10.1002/(sici)1097-0290(19970120)53:2<185::aid-bit8>3.0.co;2-k.

Tessier, L.-H., Chan, R.L., Keller, M., Weil, J.-H., Imbault, P., 1992. The *Euglena gracilis* rbcS gene contains introns with unusual borders. FEBS Lett. 304 (2—3), 252—255, http://dx.doi.org/10.1016/0014-5793(92)80631-p.

Tukey, R.H., Strassburg, C.P., 2000. Human UDP-glucuronosyltransferases: metabolism, expression, and disease. Annu. Rev. Pharmacol. Toxicol. 40, 581—616, http://dx.doi.org/10.1146/annurev.pharmtox.40.1.581.

van Leeuwen, F., Taylor, M.C., Mondragon, A., Moreau, H., Gibson, W., Kieft, R., Borst, P., 1998. β-D-Glucosyl-hydroxymethyluracil is a conserved DNA modification in kinetoplastid protozoans and is abundant in their telomeres. Proc. Natl. Acad. Sci. U. S. A. 955, 2366—2371, http://dx.doi.org/10.1073/pnas.95.5.2366.

van Luenen, H.G.A.M., Farris, C., Jan, S., Genest, P.-A., Tripathi, P., Velds, A., Kerkhoven, R.M., Nieuwland, M., Haydock, A., Ramasamy, G., Vainio, S., Heidebrecht, T., Perrakis, A., Pagie, L., van Steensel, B., Myler, P.J., Borst, P., 2012. Glucosylated hydroxymethyluracil DNA Base J, prevents transcriptional readthrough in *Leishmania*. Cell 1505, 909—921, http://dx.doi.org/10.1016/j.cell.2012.07.030.

Vocadlo, D.J., 2012. O-GlcNAc processing enzymes: catalytic mechanisms, substrate specificity, and enzyme regulation. Curr. Opin. Chem. Biol. 16 (5—6), 488—497, http://dx.doi.org/10.1016/j.cbpa.2012.10.021.

Zeidan, Q., Hart, G.W., 2010. The intersections between O-GlcNAcylation and phosphorylation: implications for multiple signaling pathways. J. Cell Sci. 1231, 13—22, http://dx.doi.org/10.1242/jcs.053678.