Published online 25 October 2008 *Nucleic Acids Research, 2009, Vol. 37, Database issue* ***D531–D538***
doi:10.1093/nar/gkn826

metaTIGER: a metabolic evolution resource

John W. Whitaker[1], Ivica Letunic[2], Glenn A. McConkey[1] and David R. Westhead[1,*]

[1]Institute of Molecular and Cellular Biology, Garstang Building, University of Leeds, Leeds, W. Yorks, LS2 9JT, UK and [2]EMBL Heidelberg, Meyerhofstraße 1, 69117 Heidelberg, Germany

Received June 10, 2008; Revised October 7, 2008; Accepted October 14, 2008

ABSTRACT

Metabolic networks are a subject that has received much attention, but existing web resources do not include extensive phylogenetic information. Phylogenomic approaches (phylogenetics on a genomic scale) have been shown to be effective in the study of evolution and processes like horizontal gene transfer (HGT). To address the lack of phylogenomic information relating to eukaryotic metabolism, metaTIGER (www.bioinformatics.leeds.ac.uk/metatiger) has been created, using genomic information from 121 eukaryotes and 404 prokaryotes and sensitive sequence search techniques to predict the presence of metabolic enzymes. These enzyme sequences were used to create a comprehensive database of 2257 maximum-likelihood phylogenetic trees, some containing over 500 organisms. The trees can be viewed using iTOL, an advanced interactive tree viewer, enabling straightforward interpretation of large trees. Complex high-throughput tree analysis is also available through user-defined queries, allowing the rapid identification of trees of interest, e.g. containing putative HGT events. metaTIGER also provides novel and easy-to-use facilities for viewing and comparing the metabolic networks in different organisms via highlighted pathway images and tables. metaTIGER is demonstrated through evolutionary analysis of *Plasmodium*, including identification of genes horizontally transferred from chlamydia.

INTRODUCTION

The volume and diversity of eukaryotic sequence data have grown exponentially over the past decade, and recent advances in sequencing will ensure that this trend increases. With this data comes increased potential for comparative genomics of the eukaryotes, and studies of eukaryotic evolution. The metabolic network is the simplest biomolecular system to predict from sequence data, because many core processes have ancient origins and are conserved across all kingdoms of life. Thus, reliable identification of orthologues is easier for core metabolic enzymes than for genes involved in less conserved processes, making them attractive components for phylogenetic studies. Equally, metabolic processes remain key drug targets, particularly for eukaryotic parasites like *Plasmodium falciparum*, so that study of pathogen metabolic networks and the comparison of these with their eukaryotic hosts is an important aspect of drug target discovery. Evolutionary information is important in this process because it can indicate distance from host enzymes and the likelihood of inhibitor cross reactivity.

There are a variety of web resources enabling the study of metabolic networks. The Kyoto Encyclopedia of Genes and Genomes (KEGG) (1) offers a set of reference pathways which have been automatically annotated for 120 eukaryotes and 592 prokaryotes, (based on KEGG 44.0). Comparing the metabolic profiles of organisms in KEGG is a manual process. BioCyc (2) pathway/genome databases are made up of three tiers of annotational accuracy (based on BioCyc 11.5), and cover a total of 360 organisms with only *Escherichia coli* in tier 1 (detailed manual annotation), another 20 organisms in tier 2 (manually checked automatic annotation) and the remainder in tier 3 (automatic annotation). BioCyc provides facilities for comparing the metabolic networks of organisms, but no phylogenetic information is provided. PUMA2 (3) contains chromosomal sequence from 369 prokaryotes and 33 eukaryotes which can be compared in terms of their metabolic networks. The evolution of different protein families can be examined, although to a limited extent as the trees are produced interactively, limiting the number of sequences that can be included. Reactome (4) is an expert-annotated predominately human database which also contains some other highly annotated organisms. It offers facilities for comparing organisms, but these do not focus on evolution, and comparisons are limited to the 23 organisms covered (based on release 22.0).

None of the above-mentioned databases bring together a broad spectrum of eukaryotic organisms with the facilities to look at the evolution of their metabolic networks on a large scale. Comparison of the enzymes present in

*To whom correspondence should be addressed. Tel: +44 113 34 33116; Fax: +44 113 34 33167; Email: d.r.westhead@leeds.ac.uk

different organisms allows the build-up and loss of pathways over evolution to be observed. The construction of phylogenetic trees on a genome scale is termed 'phylogenomics' and allows the evolution of individual genes, as well as, whole genomes to be considered. In particular, it allows the extent to which horizontal gene transfer (HGT) has occurred in eukaryotes to be investigated. HGT has for sometime been recognized as an important influence on the evolution of prokaryotes (5). It is now being realized that HGT also takes place in eukaryotes, particularly involving the gain of genes from bacteria.

In this paper we present metaTIGER, a metabolic resource that focuses upon aspects of metabolism that are not addressed elsewhere. In particular, in-depth evolutionary information about enzymes is provided in the form of 2257 maximum-likelihood phylogenetic trees, some of which contain over 500 organisms and more than 100 eukaryotes. The trees can be viewed interactively with iTOL (6) which produces intelligible displays of even the largest trees. Complex high-throughput analysis of the trees can be carried out with PhyloGenie's PHAT program (7), allowing users to define their own tree queries, which are then submitted to a Beowulf cluster for processing. Additionally, metaTIGER offers facilities that permit comparisons between eukaryotic metabolic networks in a variety of formats. The metabolic enzymes within metaTIGER are predicted using SHARKhunt (8), which operates with raw nucleic acid sequence data, including unannotated/unassembled sequence, meaning that metaTIGER can offer information on organisms that are not annotated by other facilities. As SHARKhunt's predictions are based upon sensitive sequence profile comparison techniques, enzyme assertions are likely to be more specific, and highly divergent homologues are more likely to be found than would be the case for simpler BLAST-based methods.

DATABASE CONSTRUCTION

Metabolic profiles

The sequence database behind metaTIGER metabolic profile and phylogenetic trees was constructed using SHARKhunt (8). The genomic sequence of the organisms which are covered in metaTIGER was downloaded from a number of resources (9–17) (see SI 1 for complete details), and includes information on a wide variety of eukaryotes with poor metabolic characterization and levels of genome annotation. In particular, eukaryotic taxonomic coverage was broadened by using expressed sequence tag data from the TBestDB (18). SHARKhunt scans the sequences with PSI-BLAST (19) and hidden Markov models, looking for the presence of enzyme sequence profiles that were obtained from PRIAM (20). SHARKhunt was updated to use the latest version of PRIAM which contains 2908 profiles for 2192 different E.C. (Enzyme Commission) numbers. Each profile hit is assigned an E-value. The results are then stored in the metaTIGER database.

Phylogenetic trees

Owing to the diverse taxonomic range of organisms sampled in metaTIGER, phylogenetic trees with a broad taxonomic sample could be produced. This broad taxonomic range increases the potential for new insight to be gained from exploration of the tree data. For each of the enzyme profiles a phylogenetic tree is produced from the amino acid sequences of the hits. It is advantageous to use profile-hit sequences rather than whole genes, as a hit is made of the conserved region of a protein and thus the proportion of the alignment that is made up of unconserved regions is reduced. The exclusion of non-conserved regions from alignments is important if an accurate phylogenetic tree is to be produced (21). To ensure that the trees are produced are of high quality, only sequences with profile match E-values $<10^{-30}$ were included in the trees. If the more than one sequence for a particular enzyme profile was beneath this cut-off then only the sequence with the lowest E-value was used in the tree reconstruction, reducing the chances of including paralogues. The sequences were aligned using MUSCLE (22) on default settings. Then the trees were produced using PhyML (23) using the evolutionary model JTT, the gamma-distribution model, four rate categories and invariant position. The JTT model was chosen as this has been found, most frequently, to be the best-fitting model during other phylogenomic reconstructions (24). The gamma parameter and the fraction of invariant positions were estimated from the data. The trees were optimized for topology, branch length and rate parameters. Each tree was subjected to 100 bootstrap replicates. To increase the speed for the larger trees an MPI version of PhyML was used (25); 12 of these larger trees had to have the number of sequences present reduced (see SI 2 for details) owing to memory issues when using the MPI version of PhyML. This pipeline resulted in the production of 2257 maximum-likelihood phylogenetic trees.

The user of the resource should note that each phylogenetic tree contains orthologous sequences for a specific E.C. number, and is, as far as possible, free of paralogues. They are not intended for the study of gene families containing paralogues with a variety of different functions. Rather, the trees allow the study of the evolutionary origin of specific metabolic functions and pathways in particular species or species groups. They are suitable for the detection of functional gain by HGT (as illustrated below), and for assessing the degree of evolutionary divergence between orthologous enzymes in different species. This latter application is indicated in drug target discovery, where good drug targets in a pathogen should be as divergent as possible from the host orthologue to ensure specificity. The exclusion of paralogues means that some of the trees are suitable for the estimation of species phylogenies, but this should be approached with care, because not all enzyme sequences contain sufficient phylogenetic signal to resolve species, particularly at deep branches. In these cases the tree should be viewed as interesting starting points for detailed manual study and verification. Another potential application of the trees, when used with the search procedures below, is to identify

sets of possible sequences to use in constructing species phylogenies by gene concatenation or consensus methods (26) where the avoidance of HGTs is of paramount importance.

DATABASE INTERFACE

Exploring metabolic networks

The metabolic networks in metaTIGER can be explored in four ways: (i) by using a simple search facility, (ii) by viewing KEGG map images [produced using the KEGG SOAP API (1)] that highlight the enzymes that are present in each organism, (iii) the enzymes present in a particular pathway can be compared between two organisms via a coloured KEGG pathway image and (iv) two or more organisms can be compared in a table format. A detailed description of metaTIGER search facilities along with other metaTIGER facilities can be found at: http://www.bioinformatics.leeds.ac.uk/metatiger/help.html. Additionally, all of the SHARKhunt metabolic profiles predictions can be downloaded from the site.

Viewing and searching phylogenetic trees

The phylogenetic trees can be viewed interactively using the web-based viewer iTOL (6). The use of an interactive tree viewer is necessary as some of the trees contain over 500 taxa and would not be clearly displayed through less sophisticated tree viewers. iTOL colours taxa labels according to kingdoms which means that events such as prokaryote to eukaryote HGTs can be easily identified on the large trees. As the trees are large and arbitrarily rooted the user can redefine the root and collapse branches to suit their needs. Images can be exported to files in a range of formats. An example tree for 4-hydroxy-3-methylbut-2-enyl diphosphate reductase is shown in Figure 2; this enzyme was identified below as a putative case of HGT into *Plasmodium* (see below).

It is not feasible to search manually each of the 2257 phylogenetic trees in metaTIGER for trees of interest such as those containing putative HGT events or those that would be suitable for concatenation to obtain a consensus tree depicting the evolution of species. To overcome these problems metaTIGER has a high-throughput tree searching facility, which allows users to submit their own custom tree queries. The tree queries make use of the phylogenetic analysis tool PHAT, which is part of the PhyloGenie package (7). PHAT has its own tree query language and employs a sophisticated re-rooting technique to ensure that clades being tested do not cross the root of the tree, which is important when asserting potential HGTs. As well as a tree query the user provides a minimum bootstrap value, and branches with bootstrap support below this are ignored during the analysis. The queries can be computationally demanding and are sent to a 440 core (Opteron/Linux) Beowulf cluster; the results are sent by e-mail to a user-specified address.

ILLUSTRATIVE ANALYSES

Organism profiles comparison

Unlike other Apicomplexans, including *P. falciparum*, *Cryptosporidium parvum* is not capable of *de novo* pyrimidine synthesis, owing to the absence of a six-enzyme pathway. To compensate for this *C. parvum* has three salvage enzymes (27), one of which is bi-functional. Figure 1 illustrates how metaTIGER pathway comparison facility can be used to identify such differences. If comparisons of more than two organisms are desired then the list comparison facility can be used (see Table 1). List comparisons allow for the rapid identification of organisms that contain all/part/none of a pathway of interest. In eukaryotes the shikimate pathway is present to varying degrees of completeness (28) which makes it a good test case for comparative genomics using metaTIGER. The shikimate pathway is known to be completely present in fungi, plants and red algae, heterokonts and the alveolates *Toxoplasma gondii* (29,30) and *Tetrahymena thermophila* (28), but absent in metazoans and *Cryptosporidium* (31), while *P. falciparum* is thought to have only the last three enzymes (32,33). Table 1 shows that metaTIGER's results agree with what is already known about the abundance of the shikimate pathway in eukaryotes. These two examples of using metaTIGER's pathway comparison facilities illustrate the quality of results that can be obtained from using metaTIGER.

Horizontal gene transfer analysis

In unicellular eukaryotes there are many examples of metabolic genes which have been acquired via HGT (30,34). For example, an ancestor of *Plasmodium* is thought to have gained a specialist organelle, the apicoplast, by secondary endosymbiosis, involving the gain of a red algal plastid that was originally acquired from a cyanobacterium. This secondary endosymbiosis event will have brought with it the opportunity for the endosymbiotic gene transfer (EGT), a type of HGT, of genes from the red alga to the ancestor of *Plasmodium*. Also, it has recently been suggested that the cyanobacterium's transition from free living to the endosymbiont that became the plant/algal chloroplast may have been aided by a chlamydial endosymbiont or parasite (35). Thus, genes of plant origin in *Plasmodium* may originate from cyanobacteria or chlamydia. To assess the number of metabolic enzymes that were acquired via EGT that are still present in *Plasmodium*, a high-throughput assessment of the genes, which are putatively, of plant, cyanobacterial and chlamydial origin in *Plasmodium* was carried out. When creating tree selection queries to identify trees of interest there are many variations of select statements that can be made. After optimization, the following select statement that was found to be the most accurate at identifying genes acquired from plants.

Tree select statement for plant to *Plasmodium* EGT tree selection

(Plasmodium{>1}) & ((Viridiplantae | Rhodophyta | Glaucocystophyceae){>1}) & !(Fungi | Fungi/Metazoa

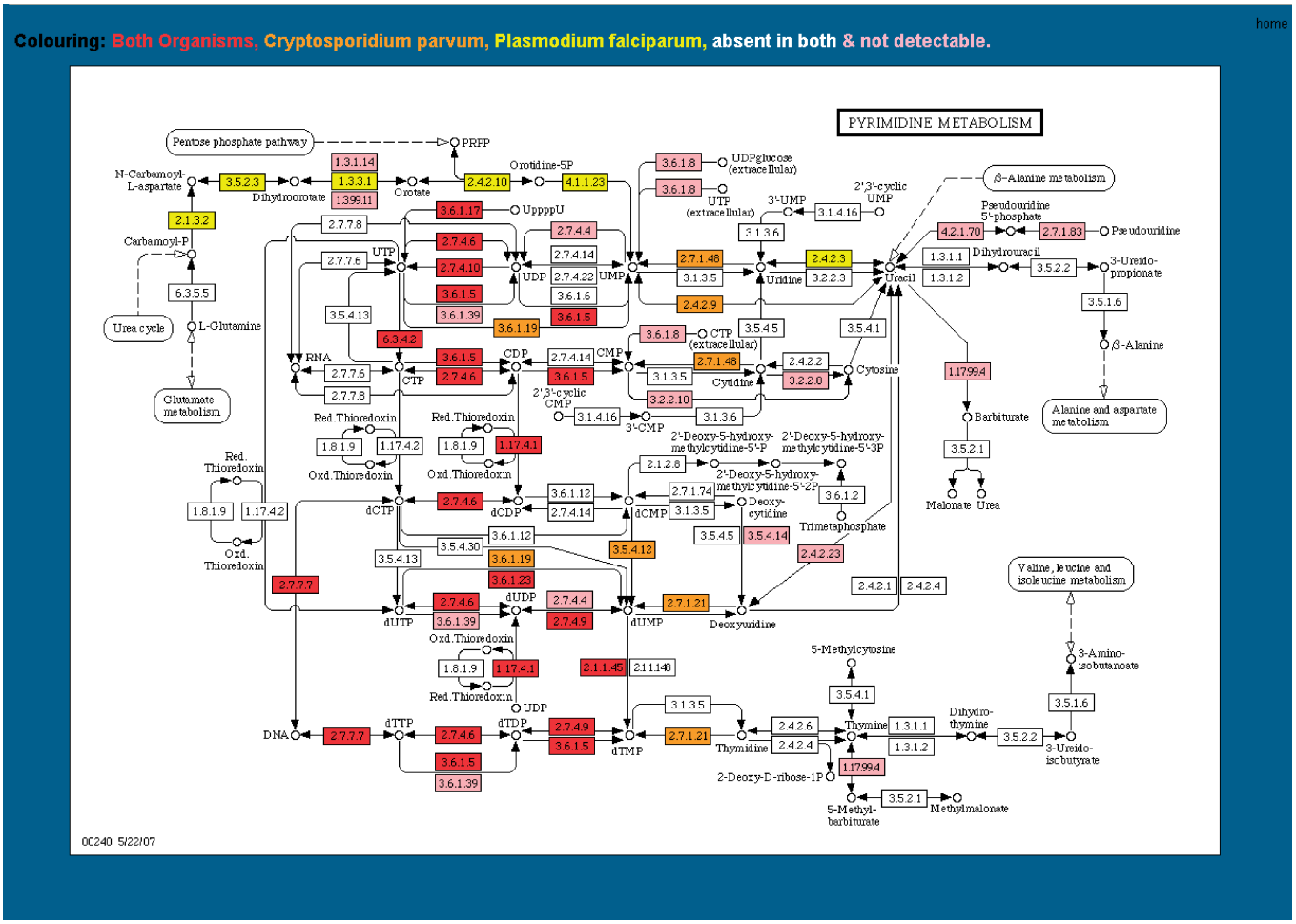

Figure 1. Screen print of pyrimidine metabolism compared between *Plasmodium falciparum* and *Cryptosporidium parvum* from the metaTIGER site. The colouring scheme is described at the top of the image as found on the webpage. Only enzymes with profile match E-values $<10^{-10}$ are included. The six-enzyme *de novo* pyrimidine biosynthesis pathway is shown in the top left image and contains the enzymes 6.3.5.5, 2.1.3.2, 3.5.2.3, 1.3.3.1, 2.4.2.10 and 4.1.1.23. Five of the enzyme in the *de novo* pyrimidine biosynthesis pathway are coloured yellow indicating that they are only found in *P. falciparum*, the sixth enzyme (6.3.5.5) is not coloured yellow as its E-value was 1.9×10^{-6} which is greater than the E-value cut-off used. Of the four pyrimidine salvage enzymes three (2.7.1.21, 2.7.1.48 and 2.4.2.9) are shown in orange as they are only found in *C. parvum* and the fourth (1.5.1.3) in is not included in this KEGG pathway image. In the image the enzyme 2.4.2.3 forms part of a pyrimidine salvage pathway and is highlighted as being present in *P. falciparum*; however, this enzyme is known to only function in pyrimidine degradation in *P. falciparum* (40,41). (NB: enzymes labelled 'not detectable' are those for which PRIAM/SHARKhunt does not have a sequence profile).

group | Pelobiontida | Malawimonadidae | Mycetozoa | Entamoebidae | Acanthamoebidae | Lobosea | Archaea) & ((Bacteria){<10})

This statement identifies clades within trees containing more than one *Plasmodium* (to avoid the possibility of bacterial contamination in a single genome leading to false identification of HGT) and plant sequences. Sequences from eukaryotes that have not ever contained plastid were not allowed because the presence of these in the clade would suggest that the gene was a general eukaryotic gene and not one gained by EGT. Selection statements that allowed one sequence from eukaryotes which have never contained a plastid were also tried: this identified 12 trees, five of which were new, although manual inspection of trees indicated that only one additional enzyme was likely to have been acquired by *Plasmodium* via EGT from plants (this extra tree was included in the analysis). Up to nine bacterial sequences were allowed because bacteria exhibit high rates of HGT and a large number of bacterial genomes (375 out of 525) are in metaTIGER. When trees were rejected on the basis of clades containing bacteria the number of trees found was reduced from seven to three. A bootstrap value of 70 was used as this has been shown to correspond to a 95% probability that the clade is real (36). Similar selection statements (shown in SI 3) were used to select trees relating to cyanobacterial and chlamydial EGT and found two and three genes, respectively. The selection statements shown in SI 3 were found to be the most accurate possible, although manual inspection of other results yielded one additional enzyme of each type. This gives a total of 11 genes: eight plant, three cyanobacteria and four chlamydia (NB. enzymes can fall in more than one group). Full details are given in Table 2 and an example is shown in Figure 2. These results were compared to the results of Huang *et al.* (37), which had previously carried out

Table 1. Comparison of the shikimate pathway between 10 eukaryotes

Enzyme name	E.C.	*Chlamydomonas reinhardtii*	*Cyanidioschyzon merolae*	*Rhizopus oryzae*	*Ciona intestinalis*	*Phytophthora ramorum*	*Toxoplasma gondii*	*Tetrahymena thermophila*	*Plasmodium falciparum*	*Plasmodium yoelii*	*Cryptosporidium parvum*
3-Deoxy-7-phosphoheptulonate synthase	2.5.1.54	1.0×10^{-125}	2.0×10^{-153}	4.0×10^{-166}		2.0×10^{-155}	1.5×10^{-83}	3.1×10^{-58}			
3-Dehydroquinate synthase	4.2.3.4	5.0×10^{-139}	7.0×10^{-103}	4.0×10^{-110}		5.0×10^{-102}	1.7×10^{-74}	7.8×10^{-95}			
3-Dehydroquinate dehydratase	4.2.1.10	1.8×10^{-41}	1.3×10^{-36}	1.9×10^{-42}	0.038	6.3×10^{-39}	8.9×10^{-15}	4.5×10^{-05}	0.006	0.001	
Shikimate dehydrogenase	1.1.1.25	9.8×10^{-59}	2.2×10^{-62}	1.6×10^{-62}	0.00033	6.4×10^{-54}	6.0×10^{-44}	2.1×10^{-16}			0.00032
Shikimate kinase	2.7.1.71	4.9×10^{-46}	2.7×10^{-31}	1.2×10^{-36}	7.8×10^{-08}	8.3×10^{-32}	8.6×10^{-20}	1.8×10^{-07}	3.3×10^{-06}	1.6×10^{-07}	8.3×10^{-07}
3-Phosphoshikimate 1-carboxyvinyltransferase	2.5.1.19	6.0×10^{-143}	5.0×10^{-130}	9.0×10^{-132}		9.0×10^{-132}	1.1×10^{-72}	3.2×10^{-42}	2.8×10^{-09}	7.7×10^{-10}	
Chorismate synthase	4.2.3.5	8.0×10^{-151}	6.0×10^{-133}	7.0×10^{-143}		2.0×10^{-134}	4.0×10^{-136}	3.0×10^{-129}	3.0×10^{-119}	8.0×10^{-118}	

The table was produced by using the metaTIGER compare lists facility to compare a custom list of enzymes that form the shikimate pathway. The profile-match E-values for the corresponding organism and E.C. number are given. The hits that have E-values $\leq 10^{-10}$ have a grey background.

HGT analysis in *P. falciparum*. We found that four out the 10 plant and cyanobacteria predictions were in common. It is not surprising that not all predictions are common as only 28% of the *P. falciparum* genes that Huang *et al.* analysed made it to the stage of phylogenetic analysis. To gain an idea of an upper limit on the number of enzymes that have been potentially gained by EGT, as some may have been missed due to the bootstrap cut-off, tree selection was carried out using no bootstrap cut-off. This found a total of 29 enzymes: 25 plant, six cyanobacteria and eight chlamydia. These results show evidence of genes with all three origins and the transport of genes into *Plasmodium* from cynaobacteria and chlamydia via the plants. Earlier studies have not included chlamydia in similar analyses or covered more than one *Plasmodium* species, but Gardner and co-workers (38) found some 30 *P. falciparum* genes (not restricted to metabolic enzymes) of probable plastid origin, consistent with our results, and with an over-representation of metabolic enzymes in EGT genes.

The enzyme (1.17.1.2) of Figure 2 is illustrative of the difficulties of this type of analysis. This enzyme is part of the non-mevalonate isoprenoid biosynthesis pathway, known to be localized in apicoplast of *Plasmodium* (39), and a case of a pathway that should be confidently of plastid evolutionary origin. On this tree 13 apicomplexans are located as a sister clade to 11 chlamydiales with 98% bootstrap support, but 12 members of plantae, two diatoms and two protozoan alga are located in another clade of 18 cyanobacteria with 70% bootstrap support. This suggests an EGT origin of this gene in plants, with the possibility of a later orthologous replacement of the gene in the Apicomplexa by HGT from chlamydials, although it is possible that this effect may be caused by low taxon sampling in the plantae. It is equally revealing that only two of the genes in this pathway have been identified as of plastid origin in our analysis. Inspection of the trees from other genes in this pathway indicates that four are potential EGTs that were excluded by our query because of bootstrap support below our stringent threshold. With this observation, our suggestion that up to 29 *Plasmodium* metabolic enzymes may have EGT origin, based on a query with no bootstrap support threshold, seems reasonable. Interestingly, the tree for the one remaining enzyme on this pathway (1-deoxy-D-xylulose-5-phosphate reductoisomerase: 1.1.1.267) indicates another possible orthologous replacement, this time from a bacterium belonging to the order rickettsiales (see SI 4 for the tree).

CONCLUDING REMARKS

We have presented a new resource for examining the predicted metabolic networks of a large number of organisms, including more than 100 eukaryotes. Because our automated enzyme annotation software is dependent only on nucleic acid sequences and not on the existence of accurate gene models and predicted protein-coding sequences, the resource is able to cover new genome sequences prior to detailed annotation. This will be a

Table 2. Summary of *Plasmodium* EGT enzymes

E.C. number	Enzyme name	Enzyme origin	KEGG pathways
2.7.1.40	Pyruvate kinase	Plant	Glycolysis/gluconeogenesis Purinemetabolism Pyruvate metabolism
5.3.1.9	Glucose-6-phosphate isomerase	Plant	Glycolysis/gluconeogenesis Pentose phosphatepathway Starch and sucrose metabolism
4.1.1.31	Phosphoenolpyruvate carboxylase	Plant	Pyruvate metabolism
1.17.4.3	4-Hydroxy-3-methylbut-2-en-1-yl diphosphate synthase	Plant	Biosynthesis of steroids
1.2.4.1	Pyruvate dehydrogenase (acetyl-transferring)	Cyanobacteria	Glycolysis/Gluconeogenesis Alanine andaspartate metabolism Valine, leucine andisoleucine biosynthesis Pyruvate metabolism
3.4.11.1	Leucyl aminopeptidase	Cyanobacteria	n/a
1.17.1.2	4-Hydroxy-3-methylbut-2-enyl diphosphate reductase	Chlamydia	Biosynthesis of steroids
2.3.1.41	Beta-ketoacyl-acyl-carrier-protein synthase I	Plant/cyanobacteria	Fatty acid biosynthesis
2.3.1.15	Glycerol-3-phosphate O-acyltransferase	Plant/chlamydia	Glycerolipid metabolism Glycerophospholipid metabolism
2.7.1.90	Diphosphate-fructose-6-phosphate 1-phosphotransferase	Plant/chlamydia	Fructose and mannose metabolism
1.3.1.9	Enoyl-[acyl-carrier-protein] reductase (NADH)	Plant/chlamydia	Fatty acid biosynthesis

The E.C. number, name, origin of the enzymes that metaTIGERs has predicted as putatively being acquired via EGT are shown. Additionally, a list of the KEGG metabolic pathways which the enzymes function within is given.

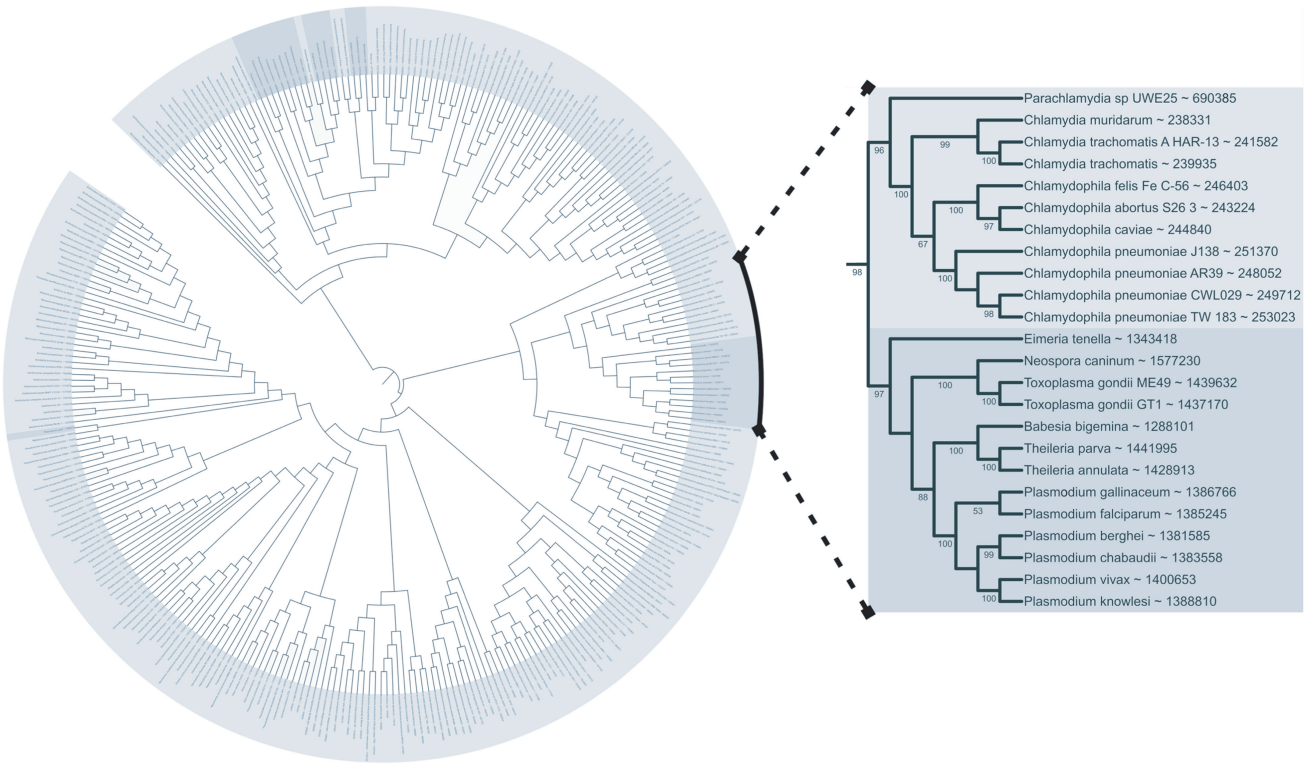

Figure 2. The metaTIGER phylogenetic tree of 4-hydroxy-3-methylbut-2-enyl diphosphate reductase (E.C. = 1.17.1.2). The tree on the right shows the entire metaTIGER phylogenetic tree as viewed on the WWW site in iTOL. The locations of bacterial leaves are highlighted with light grey and the location of eukaryotic leaves is highlighted with dark grey. The tree on the left focuses in on the location of the Apicomplexans and their chlamydial sister clade. Bootstrap values >50 are shown on the left-hand tree. For clarity branch length were ignored on both trees.

significant advantage, given the expected increases in eukaryotic sequence production in the next few years. The sensitive method of conserved motif prediction, predicts proteins of high divergence, which can be used by experimental scientists to identify proteins not in current annotations. A significant addition to the capability of existing resources is the extensive phylogenetic information that is available in the form of trees generated for enzyme coding genes, the ability to visualize these with state of the art methods, and to query them to

identify trees of interest and enzymes suitable for use in phylogenetic tree generation from multiple genes. The resource also provides highly convenient facilities for comparison of networks across multiple organisms. Currently, enzyme predictions are based on the most up-to-date set of high-quality enzyme-sequence profiles from PRIAM (20), but a future improvement will be to add to these profiles enzymes as yet without full E.C. numbers, or for which profiles are absent from PRIAM for other reasons.

ACKNOWLEDGEMENTS

Comparative genomics of this scale would not be possible without the large quantities of genomic data that are currently publicly available and for this reason the authors would like to thank all the sequencing centres and sequence repositories from which data used in this study were gained. In particular, the DOE the US Department of Energy Joint Genome Institute http://www.jgi.doe.gov/ for many genome sequences. Thanks to KEGG for their repository of reference metabolic pathways. Thanks to John W. Pinney for creating the SHARKhunt software which made it possible to conduct such large-scale genomic analysis. Thanks to Tancred Frickey for advising us on the use of PHAT. Thanks to Simon Kenworthy for helping with the aesthetics of the metaTIGER site. Finally, we thank the editor and three anonymous reviewers whose inputs have led to significant improvements in this work.

FUNDING

The BBSRC; BBSRC Research Development Fellowship (BB/C52101X/1 to D.R.W.). Funding for open access charge: University of Leeds.

Conflict of interest statement. None declared.

REFERENCES

1. Kanehisa,M., Goto,S., Hattori,M., Aoki-Kinoshita,K.F., Itoh,M., Kawashima,S., Katayama,T., Araki,M. and Hirakawa,M. (2006) From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res.*, **34**, D354–D357.
2. Caspi,R., Foerster,H., Fulcher,C.A., Kaipa,P., Krummenacker,M., Latendresse,M., Paley,S., Rhee,S.Y., Shearer,A.G., Tissier,C. *et al.* (2008) The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Res.*, **36**, D623–D631.
3. Maltsev,N., Glass,E., Sulakhe,D., Rodriguez,A., Syed,M.H., Bompada,T., Zhang,Y. and D'Souza,M. (2006) PUMA2–grid-based high-throughput analysis of genomes and metabolic pathways. *Nucleic Acids Res.*, **34**, D369–D372.
4. Joshi-Tope,G., Gillespie,M., Vastrik,I., D'Eustachio,P., Schmidt,E., de Bono,B., Jassal,B., Gopinath,G.R., Wu,G.R., Matthews,L. *et al.* (2005) Reactome: a knowledgebase of biological pathways. *Nucleic Acids Res.*, **33**, D428–D432.
5. Koonin,E.V., Makarova,K.S. and Aravind,L. (2001) Horizontal gene transfer in prokaryotes: quantification and classification. *Annu. Rev. Microbiol.*, **55**, 709–742.
6. Letunic,I. and Bork,P. (2007) Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics*, **23**, 127–128.
7. Frickey,T. and Lupas,A.N. (2004) PhyloGenie: automated phylome generation and analysis. *Nucleic Acids Res.*, **32**, 5231–5238.
8. Pinney,J.W., Shirley,M.W., McConkey,G.A. and Westhead,D.R. (2005) metaSHARK: software for automated metabolic network prediction from DNA sequence and its application to the genomes of *Plasmodium falciparum* and *Eimeria tenella*. *Nucleic Acids Res.*, **33**, 1399–1409.
9. Aguero,F., Zheng,W., Weatherly,D.B., Mendes,P. and Kissinger,J.C. (2006) TcruziDB: an integrated, post-genomics community resource for *Trypanosoma cruzi*. *Nucleic Acids Res.*, **34**, D428–D431.
10. Arnaud,M.B., Costanzo,M.C., Skrzypek,M.S., Shah,P., Binkley,G., Lane,C., Miyasato,S.R. and Sherlock,G. (2007) Sequence resources at the *Candida* Genome Database. *Nucleic Acids Res.*, **35**, D452–D456.
11. Bahl,A., Brunk,B., Crabtree,J., Fraunholz,M.J., Gajria,B., Grant,G.R., Ginsburg,H., Gupta,D., Kissinger,J.C., Labo,P. *et al.* (2003) PlasmoDB: the *Plasmodium* genome resource. A database integrating experimental and computational data. *Nucleic Acids Res.*, **31**, 212–215.
12. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Wheeler,D.L. (2008) GenBank. *Nucleic Acids Res.*, **36**, D25–D30.
13. Chisholm,R.L., Gaudet,P., Just,E.M., Pilcher,K.E., Fey,P., Merchant,S.N. and Kibbe,W.A. (2006) dictyBase, the model organism database for *Dictyostelium discoideum*. *Nucleic Acids Res.*, **34**, D423–D427.
14. Gajria,B., Bahl,A., Brestelli,J., Dommer,J., Fischer,S., Gao,X., Heiges,M., Iodice,J., Kissinger,J.C., Mackey,A.J. *et al.* (2008) ToxoDB: an integrated *Toxoplasma gondii* database resource. *Nucleic Acids Res.*, **36**, D553–D556.
15. Heiges,M., Wang,H., Robinson,E., Aurrecoechea,C., Gao,X., Kaluskar,N., Rhodes,P., Wang,S., He,C.-Z., Su,Y. *et al.* (2006) CryptoDB: a *Cryptosporidium* bioinformatics resource update. *Nucleic Acids Res.*, **34**, D419–D422.
16. Pruitt,K.D., Tatusova,T. and Maglott,D.R. (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.*, **35**, D61–D65.
17. Sherman,D., Durrens,P., Iragne,F., Beyne,E., Nikolski,M. and Souciet,J.-L. (2006) Genolevures complete genomes provide data and tools for comparative genomics of hemiascomycetous yeasts. *Nucleic Acids Res.*, **34**, D432–D435.
18. O'Brien,E.A., Koski,L.B., Zhang,Y., Yang,L., Wang,E., Gray,M.W., Burger,G. and Lang,B.F. (2007) TBestDB: a taxonomically broad database of expressed sequence tags (ESTs). *Nucleic Acids Res.*, **35**, D445–D451.
19. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
20. Claudel-Renard,C., Chevalet,C., Faraut,T. and Kahn,D. (2003) Enzyme-specific profiles for genome annotation: PRIAM. *Nucleic Acids Res.*, **31**, 6633–6639.
21. Talavera,G. and Castresana,J. (2007) Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Syst. Biol.*, **56**, 564–577.
22. Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.
23. Guindon,S. and Gascuel,O. (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.*, **52**, 696–704.
24. Huerta-Cepas,J., Dopazo,H., Dopazo,J. and Gabaldon,T. (2007) The human phylome. *Genome Biol.*, **8**, R109.
25. Torres,M., Vieira,C., Gonçalves,G. and Junior,Z. (2007), *Brazilian Symposium on Bioinformatics 2007*, Brazil, pp. 115–127.
26. Ciccarelli,F.D., Doerks,T., von Mering,C., Creevey,C.J., Snel,B. and Bork,P. (2006) Toward automatic reconstruction of a highly resolved tree of life. *Science*, **311**, 1283–1287.
27. Striepen,B., Pruijssers,A.J., Huang,J., Li,C., Gubbels,M.J., Umejiego,N.N., Hedstrom,L. and Kissinger,J.C. (2004) Gene transfer in the evolution of parasite nucleotide biosynthesis. *Proc. Natl Acad. Sci. USA*, **101**, 3154–3159.

28. Richards,T.A., Dacks,J.B., Campbell,S.A., Blanchard,J.L., Foster,P.G., McLeod,R. and Roberts,C.W. (2006) Evolutionary origins of the eukaryotic shikimate pathway: gene fusions, horizontal gene transfer, and endosymbiotic replacements. *Eukaryot. Cell*, **5**, 1517–1531.
29. Campbell,S.A., Richards,T.A., Mui,E.J., Samuel,B.U., Coggins,J.R., McLeod,R. and Roberts,C.W. (2004) A complete shikimate pathway in *Toxoplasma gondii*: an ancient eukaryotic innovation. *Int. J. Parasitol.*, **34**, 5–13.
30. Nosenko,T. and Bhattacharya,D. (2007) Horizontal gene transfer in chromalveolates. *BMC Evol. Biol.*, **7**, 173.
31. Abrahamsen,M.S., Templeton,T.J., Enomoto,S., Abrahante,J.E., Zhu,G., Lancto,C.A., Deng,M., Liu,C., Widmer,G., Tzipori,S. *et al.* (2004) Complete genome sequence of the apicomplexan, *Cryptosporidium parvum*. *Science*, **304**, 441–445.
32. McConkey,G.A., Pinney,J.W., Westhead,D.R., Plueckhahn,K., Fitzpatrick,T.B., Macheroux,P. and Kappes,B. (2004) Annotating the *Plasmodium* genome and the enigma of the shikimate pathway. *Trends Parasitol.*, **20**, 60–65.
33. McRobert,L., Jiang,S., Stead,A. and McConkey,G.A. (2005) *Plasmodium falciparum*: Interaction of shikimate analogues with antimalarial drugs. *Exp. Parasitol.*, **111**, 178.
34. Huang,J., Mullapudi,N., Lancto,C.A., Scott,M., Abrahamsen,M.S. and Kissinger,J.C. (2004) Phylogenomic evidence supports past endosymbiosis, intracellular and horizontal gene transfer in *Cryptosporidium parvum*. *Genome Biol.*, **5**, R88.
35. Huang,J. and Gogarten,J.P. (2007) Did an ancient chlamydial endosymbiosis facilitate the establishment of primary plastids? *Genome Biol.*, **8**, R99.
36. Hillis,D.M. and Bull,J.J. (1993) An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. *Syst. Biol.*, **42**, 182.
37. Huang,J., Mullapudi,N., Sicheritz-Ponten,T. and Kissinger,J.C. (2004) A first glimpse into the pattern and scale of gene transfer in Apicomplexa. *Int. J. Parasitol.*, **34**, 265–274.
38. Gardner,M.J., Hall,N., Fung,E., White,O., Berriman,M., Hyman,R.W., Carlton,J.M., Pain,A., Nelson,K.E., Bowman,S. *et al.* (2002) Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature*, **419**, 498.
39. Roos,D.S., Crawford,M.J., Donald,R.G., Fraunholz,M., Harb,O.S., He,C.Y., Kissinger,J.C., Shaw,M.K. and Striepen,B. (2002) Mining the *Plasmodium* genome database to define organellar function: what does the apicoplast do? *Philos. Trans. R Soc. Lond. B Biol. Sci.*, **357**, 35–46.
40. Reyes,P., Rathod,P.K., Sanchez,D.J., Mrema,J.E., Rieckmann,K.H. and Heidrich,H.G. (1982) Enzymes of purine and pyrimidine metabolism from the human malaria parasite, *Plasmodium falciparum*. *Mol. Biochem. Parasitol.*, **5**, 275–290.
41. Walsh,C.J. and Sherman,I.W. (1968) Isolation, characterization and synthesis of DNA from a malaria parasite. *J. Protozool.*, **15**, 503–508.