Published in Molecular Phylogenetics and Evolution 47, issue 1, 366-377, 2008
which should be used for any reference to this work

Lateral transfer of the gene for a widely used marker, α-tubulin, indicated by a multi-protein study of the phylogenetic position of *Andalucia* (Excavata)

Alastair G.B. Simpson [a,b,*], Thomas A. Perley [a], Enrique Lara [c]

[a] *Department of Biology, Dalhousie University, Halifax, Nova Scotia B3H 4J1, Canada*
[b] *Canadian Institute for Advanced Research, Program in Integrated Microbial Biodiversity*
[c] *Unité d'Ecologie, Systématique et Evolution—CNRS UMR8079 Université Paris-Sud 11, Bâtiment 360, 91405 Orsay Cedex, France*

Abstract

α-Tubulin is one of the most widely used markers for estimating deep-level phylogenetic relationships amongst eukaryotes. We sequenced 6–7 nuclear protein-coding genes, including α-tubulin, from the two described species of the enigmatic jakobid(-like) excavate protist *Andalucia*. Concatenated protein phylogenies place *Andalucia* in a clade with other jakobids, Euglenozoa and Heterolobosea. Individual gene trees, except that of α-tubulin, do not conflict strongly with this position. In α-tubulin trees, *Andalucia* instead falls in a strongly supported clade with diplomonads, parabasalids and opisthokonts (including animals and fungi), and branches with diplomonads. This clade is robust to changes in taxon sampling, and is unlikely to represent long-branch attraction, compositional heterogeneity artefact, or segmental gene conversion. Phylogenies estimated without α-tubulin strongly support the original position for *Andalucia*, and also reinforce recent studies in placing diplomonads and parabasalids with Preaxostyla, not opisthokonts. α-Tubulin seems to have experienced two or more eukaryote-to-eukaryote lateral gene transfer (LGT) events, one perhaps from an ancestral opisthokont to an ancestor of diplomonads and parabasalids, or vice versa, and one probably from the diplomonad lineage to *Andalucia*. Like EF-1α/EFL, α-tubulin has a complex history that needs to be taken into account when using this marker for deep-level phylogenetic inference.

Keywords: Excavate; Eukaryote evolution; Horizontal gene transfer; Jakobid; Lateral gene transfer; Opisthokont; Protist; Protozoa; *Reclinomonas*; Tubulin

1. Introduction

Phylogenetic inference from nucleus-encoded proteins serves an important role in improving our understanding of the eukaryotic portion of the tree of life, alongside estimates based on nuclear ribosomal RNA sequences and organellar genes. Nucleus-encoded protein phylogenies have been crucial in identifying and/or strongly confirming many high-level groupings amongst eukaryotes that are now widely, if not universally accepted (see Adl et al., 2005; Baldauf, 2003; Keeling et al., 2005; Parfrey et al., 2006; Simpson and Roger, 2004). These include: opisthokonts, the super-group that contains animals and fungi (Baldauf and Palmer, 1993), stramenopiles and alveolates (Harper et al., 2005; Simpson et al., 2006), Microsporidia with or within Fungi (Edlind et al., 1996; Keeling and Doolittle, 1996; Li et al., 1996), Euglenozoa and Heterolobosea (Baldauf et al., 2000), Rhizaria, or at least a Foraminifera–Cercozoa relationship (Keeling, 2001), Amoebozoa (Baldauf et al., 2000; Bapteste et al., 2002) and Amoebozoa with opisthokonts (Baldauf et al., 2000). Recently, phylogenies of datasets largely or entirely composed of nucleus-encoded proteins have been used to argue strongly that jakobids are related to Euglenozoa and Heterolobosea

* Corresponding author. Address: Dalhousie University, Department of Biology, 1355 Oxford St, Halifax, Nova Scotia, B3H 4J1, Canada. Fax: +1 902 494 3736.
E-mail address: alastair.simpson@dal.ca (A.G.B. Simpson).

(Rodríguez-Ezpeleta et al., 2007; Simpson et al., 2006), that metamonads *sensu* Cavalier-Smith (2003) are a clade (Hampl et al., 2005), and that Rhizaria is related to stramenopiles and alveolates (Burki et al., 2007; Hackett et al., 2007; Rodríguez-Ezpeleta et al., 2007).

Studies based largely or entirely on nucleus-encoded proteins differ widely in the quality and quantity of data. At one extreme, many studies employ a few (typically 1–10) proteins, usually chosen deliberately because they are slowly evolving, essential and well-sampled (recent examples: Hampl et al., 2005; Harper et al., 2005; Kim et al., 2006; Shalchian-Tabrizi et al., 2006; Simpson et al., 2006; Steenkamp et al., 2006). The proteins most commonly examined include actin, α-tubulin, β-tubulin, elongation factor 1α (EF-1α), elongation factor 2 (EF-2), cytosolic heat shock protein 70 (HSP70), cytosolic heat shock protein 90 (HSP90), and the largest subunit of RNA polymerase II. At the other extreme, phylogenomic approaches include sequences for tens or hundreds of proteins, obtained primarily from genome sequencing and EST projects (Bapteste et al., 2002; Burki and Pawlowski, 2006; Burki et al., 2007; Patron et al., 2007; Philippe et al., 2004; Rodríguez-Ezpeleta et al., 2005, 2007). Phylogenomic datasets, however, typically include many rapidly evolving and/or sparsely distributed proteins. Studies employing intermediate quantities of data also exist, thanks to unusually extensive PCR-based efforts, small-scale EST projects on the taxa of interest, or extraction of the 'best' genes from large phylogenomic datasets (Arisue et al., 2005; Ciccarelli et al., 2006; Hackett et al., 2007; Nozaki et al., 2007).

Phylogenetic approaches that use slowly evolving, essential proteins have several potential virtues. All things being equal, slowly evolving sequences will preserve historical signal over more time than rapidly evolving sequences. Slowly evolving proteins typically exhibit little length variation, and thus are easy and uncontroversial to align, eliminating a potentially significant source of ambiguity that can affect analyses of rapidly evolving proteins (and ribosomal RNA genes). Further, it is believed by some that genes for essential proteins with core functions are more refractory to gene transfer between species than are sparsely distributed genes, and thus the gene's history is more likely to faithfully reflect organismal phylogeny (see Andersson, 2005).

Nonetheless recent studies have shown that some of the nuclear protein-coding genes used most widely for deep-level phylogenetic inference have more complex evolutionary histories than first realized, even after excluding endosymbiotic gene replacement events, and cases likely to represent phylogenetic analysis artefact. Most spectacularly, the normal eukaryotic form of elongation factor 1α (EF-1α) is not detected in certain taxa scattered across the breadth of the eukaryotic tree. These organisms instead have an 'elongation factor-like' protein (EFL) that is related to EF-1α, but is very different in sequence (Gile et al., 2006; Keeling and Inagaki, 2004; Noble et al., 2007). EFL has apparently 'replaced' EF-1α many times (e.g., Noble et al., 2007), and EFL genes have probably been transferred laterally between eukaryotes on several occasions (Keeling and Inagaki, 2004). From the perspective of reconstructing organismal phylogeny, it is fortunate that eukaryotic EF-1α and EFL are easily distinguishable, and form separate clades in phylogenetic trees (Keeling and Inagaki, 2004). For this reason, recent molecular phylogenetic studies that include EF-1α can and do deliberately exclude EFL sequences (e.g., Harper et al., 2005; Simpson et al., 2006; Steenkamp et al., 2006).

Other studies have recovered unexpected phylogenetic patterns from α-tubulin data. Phylogenies of α-tubulin, or of α-tubulin and β-tubulin combined, almost always group sequences from parabasalids, diplomonads, and *Carpediemonas* (a close relative of diplomonads) with those from opisthokonts, with strong statistical support (Hampl et al., 2005; Simpson et al., 2002, 2006, see also Arisue et al., 2005). By contrast, other well-sampled genes only rarely recover this relationship, and never with high support. It is suspected that the strong signal in the α-tubulin data is not reflecting the organismal phylogeny (Hampl et al., 2005; Simpson et al., 2006). This is difficult to demonstrate, however, since the position of the diplomonad–parabasalid grouping itself has not been established with any confidence. Several studies of SSUrRNA genes or multiple proteins show a relationship with Preaxostyla (oxymonads and *Trimastix*), but statistical support is usually low unless taxon sampling is very limited (Berney et al., 2004; Cavalier-Smith, 2003; Hampl et al., 2005; Simpson, 2003; Simpson et al., 2002, 2006).

In this study, we consider another example of conspicuous phylogenetic aberrancy in a widely used marker sequence, again, α-tubulin. This case concerns the enigmatic taxon *Andalucia*, a clade of small heterotrophic flagellates with 'excavate-type' feeding grooves (Lara et al., 2006; Simpson, 2003). The group contains just two described species, *Andalucia incarcerata* (formerly *Jakoba incarcerata*) and *Andalucia godoyi* (Lara et al., 2006). Transmission electron microscopy studies of both species demonstrate that their cytoskeletal organizations are very similar to those of previously studied members of the excavate group Jakobida, such as *Jakoba* and *Reclinomonas* (Lara et al., 2006; Simpson and Patterson, 2001). Phylogenetic analyses of these morphological data place *Andalucia* in a strongly supported clade with (other) jakobids, relative to other excavates (Lara et al., 2006). Published SSUrRNA gene phylogenies, however, never group *Andalucia* specifically with other jakobids (Lara et al., 2006; Simpson et al., 2002). At closest, *Andalucia* and other jakobids fall as two separate branches within a larger clade that also includes Euglenozoa and Heterolobosea, and statistical support even for this larger clade is relatively weak (Berney et al., 2004; Cavalier-Smith, 2003, 2004; Lara et al., 2006; Nikolaev et al., 2004a, b), unless outgroup taxon sampling is extremely limited (Lara et al., 2007). Interestingly, studies of tubulin proteins including one of the two species (*A. incarcerata*) place this organism in a poorly resolved

position in β-tubulin phylogenies, but with or within the suspicious (diplomonads, parabasalids, opisthokonts) clade in α-tubulin trees and combined tubulin phylogenies (Edgcomb et al., 2001; Simpson et al., 2002).

Here, we examine the phylogenetic placement of *Andalucia* using a dataset of seven slowly evolving nuclear protein-coding genes, including α-tubulin, and containing data both from *A. godoyi*, and from a novel isolate of *A. incarcerata*. Our analysis, especially of sequences other than α-tubulin, strongly supports the placement of *Andalucia* with other jakobids, Euglenozoa, and Heterolobosea, thereby highlighting the discordant position of *Andalucia* in α-tubulin trees. In the absence of any signature of phylogenetic artefact, it is likely that *Andalucia* acquired its current α-tubulin gene via eukaryote-to-eukaryote LGT.

2. Methods

2.1. Organism isolation and identification; DNA extraction

Andalucia godoyi strain And28, originally described by Lara et al. (2006, 2007), was grown in 30 ml tissue culture flasks containing 6 ml of Neff's (=Page's) Amoeba Saline (AS) supplemented by 1:300 v/v LB medium (see Lara et al., 2006). Genomic DNA was extracted from a dense culture ($\sim 10^6$ cells/ml) using CTAB and organic extractions, after Clark and Diamond (1991).

Andalucia incarcerata (MB1 isolate) was isolated in October 2004 from organically enriched and sulphidic intertidal sediment in a sheltered embayment within Mahone Bay, Nova Scotia, Canada (44°21′N, 64°20′W). Samples were taken at low tide, following removal of the overlying aerobic layer. Approximately 2 ml of slurry was added to sealed 30 ml serum vials containing 10 ml total of autoclave-sterilised seawater diluted 1:1 with sterile water, and an autoclaved barley grain. Samples were incubated at 21 °C, and observed periodically. After three weeks, *Andalucia incarcerata* was observed in one vial (identification by phase contrast microscopy), and a subculture was started. Several rounds of subculturing were carried out on an *ad hoc* basis, with cultures held at 21–25 °C. The culture line was apparently monoprotistan after five transfers, and no eukaryotic contaminant has been observed since in >36 months of continuous cultivation. Cultures were subsequently maintained in 15 ml polypropylene tubes containing 12 ml of one of two media: (i) sterile seawater mixed 1:1 with AS, plus 1:30 v/v LB media; (ii) sterile seawater mixed 1:1 with Sonneborn's *Paramecium* medium (ATCC 802 media). For maintenance, cultures were kept at 21 °C and subcultured every 10–14 days. Genomic DNA was extracted as for *A. godoyi* (see above), except that two phenol–chloroform extractions were performed, not one.

To compare MB1 to the previously studied organisms, especially the original Quibray Bay (Australia) isolate of *A. incarcerata* (Bernard et al., 2000), the SSUrRNA gene sequence from MB1 was amplified by polymerase chain reaction (PCR) using the universal eukaryotic primers 'EukA' and 'EukB' (Medlin et al., 1988) according to the protocol described in Section 2.2 below, but without acetamide in the PCR cocktail. The PCR product was purified using a PCR clean-up column kit (Sigma), and directly sequenced. The SSUrRNA gene sequence was added manually to an alignment of eukaryotic sequences including those from the Quibray Bay isolate of *A. incarcerata*, from *A. godoyi* (see Lara et al., 2006) and from environmental sequences with high similarity to *Andalucia* SSUrRNA genes (e.g., Luo et al., 2005; Behnke et al., 2006). Two datasets were analysed, one containing only 'short-branching' eukaryotes, and retaining 1406 well-aligned sites, the second also containing representatives of Heterolobosea and Euglenozoa (two 'long-branching' groups), and retaining 1198 sites. Maximum Likelihood (ML) trees were estimated under a 'GTR+Γ+invariable sites' model using the 'May 2006' version of TREEFINDER (Jobb et al., 2004), with a 500 replicate bootstrap analysis performed for the former dataset under the same model.

2.2. Sequencing of protein-coding genes

Seven well-sampled nucleus-encoded marker genes were examined: actin, α-tubulin, β-tubulin, EF-1α, EF-2, HSP70 and HSP90. The majority of the coding region from each of these genes was amplified from *Andalucia godoyi* (And28 isolate) and *Andalucia incarcerata* (MB1 isolate) using degenerate PCR. Supplementary material, Table 1 provides a list of degenerate primers used, including the sequences of new designs. A list of the primers used for each successful PCR-amplification is given in Supplementary material, Table 2. All PCR-amplifications were performed in 10 or 20 μl of a cocktail containing 0.05 U/μl Taq polymerase (Sigma or Invitrogen), 1.5 mM Mg^{2+}, 0.2 mM dNTPs, 1 μM primers, and (in most cases) 5% w/v acetamide, all in 1× PCR buffer (Sigma or Invitrogen). Amplification experiments were run for 40 cycles, with annealing temperatures of 48.5–55 °C. Amplification products of the expected sizes, as determined by gel electrophoresis, were gel-extracted, and ligated into a TA plasmid vector (TOPO 2.1; Invitrogen), and used to transform competent *Escherichia coli* (TOP10; Invitrogen). Positive colonies were PCR-screened to confirm insert size, and one or more positive clones (usually four) were grown overnight in liquid LB. The plasmid DNA was extracted using a miniprep kit (Sigma) and end-sequenced. The reads were managed and compared using Sequencher version 4.2 (Gencodes Corp.), and amplicon identities were established by BLASTX searches against the GenBank protein database. In most cases one clone was then selected for complete sequencing, using primer walking where required. Two distinct paralogs of *Andalucia godoyi* α-tubulin were encountered, and one clone of each was completely sequenced. Attempts to amplify HSP70 from *Andalucia incarcerata* were unsuccessful. New sequences have the Genbank Accession Nos. EU334873–EU334887.

Sequences were conceptually translated to amino acids for subsequent analyses.

2.3. Phylogenetic analysis

The amino acid sequences from *A. godoyi* and *A. incarcerata* were aligned by eye to putatively orthologous sequences from other eukaryotes. For the main analyses 47 taxa were selected that represented broadly the major groups of eukaryotes and also included >30% of well-aligned sites from at least six of the seven examined proteins, >50% of well-aligned sites from at least five of the proteins and >70% of well-aligned sites for at least four. Most of these taxa were individual species, but in some cases data from two related species (usually from the same genus) were combined to represent one OTU. Gene families, especially those within animals or land plants, were examined to reduce potential paralog-sampling problems, as in Simpson et al. (2006). After exclusion of ambiguously aligned positions, 3597 sites were retained for analysis. The resulting data matrix was 90% populated (89% when α-tubulin was excluded—see below), with the least covered taxon (*Corallochytrium*) still represented at 62% of sites (59% without α-tubulin). *Andalucia godoyi* and *A. incarcerata* were represented at 87% and 74% of included sites, respectively (87% and 72% without α-tubulin).

We also examined a dataset based on the same alignment that included additional taxa of phylogenetic importance, including some that were less well-sampled than allowed in our main dataset (e.g., apusomonads and cryptophytes). This second dataset contained 54 taxa, was 89% populated (88% without α-tubulin), and the least well-sampled taxon (the apusomonad *Apusomonas*) was represented at 51% of sites (47% without α-tubulin). Alignments are available by request to AGBS.

Phylogenetic trees were estimated using the Whelan and Goldman (WAG) amino acid substitution matrix, with among-site rate variation modelled by a Γ distribution. Initially, we examined the data using a 'concatenated' approach (i.e., treating all the data as a single 'super-gene'). In this approach, tree-searching was performed in a ML framework using two programs: (i) RAxML v2.2.3 (Stamatakis, 2006), with 50 replicates, each beginning from a parsimony tree constructed by random taxon addition, and with the Γ distribution approximated by four discrete categories; (ii) PHYML v2.4.5 (Guindon and Gascuel, 2003) using an eight discrete category approximation of a Γ distribution, with the shape parameter α estimated previously under the neighbour-joining tree by TREE-PUZZLE 5.2 (Strimmer and von Haeseler, 1996). We performed a Bayesian analysis using MRBAYES v3.1.2 (Ronquist and Huelsenbeck, 2003) under the same model, except that branch lengths and α parameter were estimated separately for each gene ('unlinked' model). Default priors were used. Two independent analyses were run with three heated and one cold chain (temperature parameter 0.2) for 1.93×10^6 generations, with a sampling frequency of 0.01 and a conservative 1.4×10^6 generations discarded as burn-in following examination of the likelihood plots (note that the two runs had stabilised, but had not converged). The ML tree was selected by comparing all unique tree topologies recovered in the RAxML replicates (10), the PHYML analysis (1), and all tree topologies sampled in either run of the bayesian analysis (7) under both our concatenated and unlinked models, using TREE-PUZZLE as the common platform for likelihood estimation (Γ distribution approximated by eight discrete categories).

Support for the recovered topologies was evaluated using bootstrapping under the concatenated model. A 500 replicate bootstrap analysis was performed using PHYML under the WAG+Γ model described above. As a check for the effectiveness of tree searching in PHYML, we also performed a 200 replicate bootstrap analysis in RAxML, with the WAG substitution matrix, and the 'CAT' approximation to model among-site rate variation, with 25 categories (parsimony-based starting trees, as described above).

We more briefly examined the 54-taxon dataset. PHYML and RAxML were used to search for the ML tree under the concatenated model as described above, except that the Γ distribution was approximated by four discrete categories. Bootstrap support under the same model was assessed using PHYML, with 200 re-sampling replicates.

The searches for the best tree and bootstrap support analyses were repeated with the α-tubulin data excluded (3173 sites remaining), as described above, except that the Bayesian analysis used four independent runs of 1×10^6 generations, with 7×10^5 generations burn-in (none of the runs converged). For the 47-taxon dataset, 12 unique tree topologies were recovered in the RAxML replicates, and 26 were sampled in the Bayesian analyses.

All the gene alignments were also analysed individually. Under a WAG+Γ model (four categories) we searched for the ML tree for each dataset using RAxML and PHYML and assessed bootstrap support using PHYML (200 replicates). In all cases, the ML tree estimated using RAxML had a higher likelihood than the tree found using PHYML when compared on a common platform (PHYML). In cases where the alignments contained sequences with <50% of sites represented (actin, β-tubulin, EF-2 and HSP90) the analyses were repeated with these sequences excluded. In the case of α-tubulin, the bootstrap analysis was redone as a 1000 replicate analysis (with an eight category Γ approximation)—there was little relevant difference in the results.

We performed a suite of additional phylogenetic analyses of the α-tubulin data alone. Firstly we constructed a second α-tubulin dataset in which we increased taxon sampling, but excluded highly divergent sequences. Notable additions included *Telonema* spp., kathablepharids and *Apusomonas*. The inclusion of *Apusomonas* is significant because it branches strongly with opisthokonts in α-tubulin trees (Kim et al., 2006). Notable exclusions were non-flagellated Amoebozoa, *Sawyeria*, red algae and fungi other

than chytrids. This second dataset included 75 taxa and 424 sites. The ML tree and statistical support for this dataset were estimated as for the original α-tubulin analyses (see above). For both this 75-taxon dataset, and the original 47-taxon α-tubulin dataset, amino acid composition was examined using the χ^2 test in TREE-PUZZLE. The χ^2 tests were repeated with all constant sites excluded. Branch lengths across the tree were compared using the values estimated for the ML tree by RAxML.

For both α-tubulin datasets we performed ML bootstrap analyses (PHYML, WAG+Γ model, eight rate categories, 1000 replicates), excluding either opisthokonts or diplomonads and parabasalids. For the 75-taxon dataset we also ran similar analyses excluding *Apusomonas* plus either opisthokonts or diplomonads and parabasalids.

Andalucia α-tubulin shares several unusual amino acid characters with either (i) diplomonads and parabasalids, (ii) diplomonads, parabasalids and opisthokonts, or (iii) diplomonads, parabasalids, opisthokonts and Amoebozoa. Many of these are located in one short segment of the alignment, corresponding to residues 190–204 in *Homo* α-tubulin 1. To test whether this segment was responsible for the observed phylogenetic position of *Andalucia* α-tubulin, we excluded these 15 amino acids from both α-tubulin datasets and performed ML bootstrap analyses (PHYML, WAG+Γ, eight rate categories, 1000 replicates), as described above.

3. Results

3.1. Identity of the new isolate of Andalucia incarcerata

The new 'MB1' isolate of *Andalucia incarcerata* is morphologically indistinguishable from the original isolate from Quibray Bay, Australia, and was extracted from a similar habitat (sulphide-rich and oxygen-poor marine intertidal sediment). Their SSUrRNA gene sequences are 98% identical, and these form a strongly supported clade with each other and with some environmental sequences (see Supplementary material, Fig. 1). Interestingly, the related environmental sequences come from sulphide-rich water from a Norwegian fjord (Behnke et al., 2006), and a sulphide-rich spring in mid-west USA (Luo et al., 2005). The α-tubulin and β-tubulin gene sequences from MB1 are also very similar to those from the Quibray Bay isolate (96–97% identity at the DNA level; 99–100% identical in amino acid sequence). Bearing in mind the close genetic, morphological and ecological similarity to the Quibray Bay isolate (from which the type is derived), we regard MB1 as belonging to the same nomenclatural species, *Andalucia incarcerata*.

SSUrRNA gene phylogenies confirm that *Andalucia godoyi* and *A. incarcerata* are close relatives, although at least two other distinct, shallow clades represented only by environmental sequences group with the two described *Andalucia* species (Supplementary material, Fig. 1; see also Behnke et al., 2006). In the analysis including only 'short-branching' SSUrRNA gene sequences *Andalucia* branches quite strongly with other jakobids. As in previous studies however, this clade is interrupted by Heterolobosea and Euglenozoa when these 'long-branching' taxa are included (Supplementary material, Fig. 1; see also Lara et al., 2006).

3.2. Combined gene analysis

The ML trees estimated for the seven-gene, 47-taxon dataset under both concatenated and unlinked models are mostly consistent with similar previous studies of eukaryote phylogeny (Fig. 1). Opisthokonts form a strong clade, and branch close to a moderately supported Amoebozoa clade. Stramenopiles and alveolates are each strongly supported, and form a strong clade. Archaeplastida (Chloroplastida/Viridaeplantae, red algae and glaucophytes) are close to each other, but the weakly supported clade that unites them unexpectedly also includes haptophytes and the rhizarian *Bigelowiella*.

As is typical, excavates are not monophyletic in the ML trees. Diplomonads and parabasalids form a strong clade that is related to opisthokonts. Preaxostyla (*Trimastix* and oxymonads) form a strong clade that represents a separate deep branch within the tree. Jakobids, Heterolobosea and Euglenozoa form a third excavate clade that also includes *Andalucia* (this clade is described in more detail below). Finally, *Malawimonas* falls in a weakly supported position at the base of the (*Andalucia*, other jakobids, Heterolobosea, Euglenozoa) clade with the concatenated model, but as the sister group to Preaxostyla with the unlinked model.

The (*Andalucia*, other jakobids, Heterolobosea, Euglenozoa) clade receives weak PHYML bootstrap support (58%). The four groups within the clade each receive strong bootstrap support (*Andalucia*, 100%; other jakobids, 100%; Heterolobosea, 88%; Euglenozoa, 100%), however the relationships amongst these groups are poorly resolved. In the ML trees *Andalucia* branches with other jakobids, but bootstrap support is trivial (13%). Bootstrapping using RAxML gives broadly similar results, except that the support for the (*Andalucia*, other Jakobids, Heterolobosea, Euglenozoa) clade is quite strong (82%). Support for (*Andalucia*, other jakobids) is higher, but still very low (33%).

A very similar topology is recovered in the analysis (concatenated model) of the 54-taxon dataset (with cryptophytes weakly related to haptophytes, and apusomonads weakly sister to opisthokonts). The (*Andalucia*, other jakobids, Heterolobosea, Euglenozoa) clade receives 69% PHYML bootstrap support. Within this clade *Andalucia* branches as sister to other jakobids in the ML tree, but support is extremely low (8%).

3.3. Single gene analyses

When each gene is analysed individually, a striking difference is apparent in the phylogenetic signal for the place-

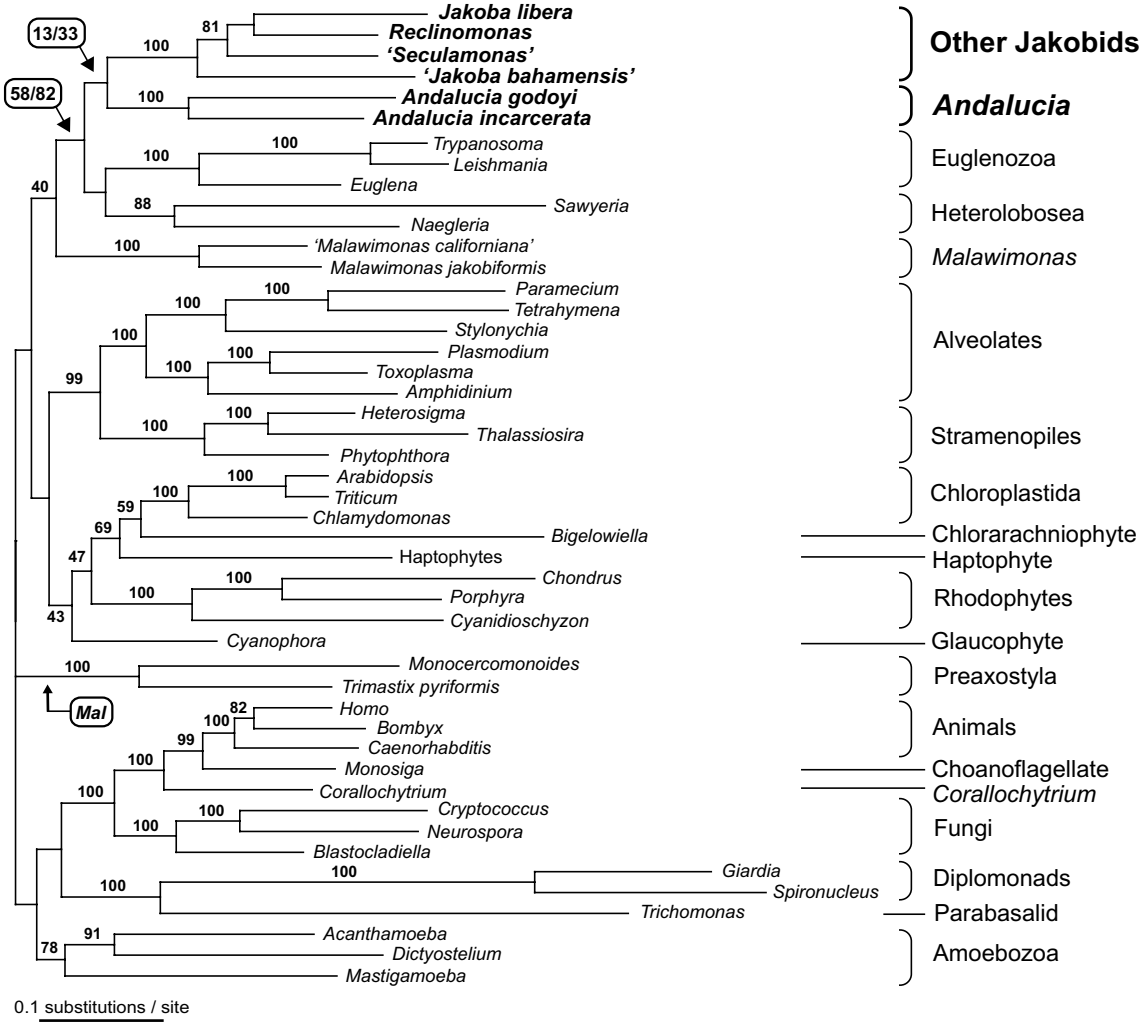

Fig. 1. Maximum likelihood tree for the full seven-protein dataset, under the concatenated ('super-gene') model (WAG+Γ). Percentage bootstrap support values are depicted. Single numbers represent values estimated using PHYML (WAG+Γ, 500 replicates). Most values <40% are not shown. For important nodes double numbers are shown giving the PHYML values followed by values estimated using RAxML (WAG+CAT, 200 replicates). The arrow labelled '*Mal*' shows the position of the *Malawimonas* clade in the maximum likelihood tree under the unlinked model (there were no other important differences between the concatenated and unlinked trees).

ment of *Andalucia* from the α-tubulin data when compared to the other proteins. Most nodes within the backbones of the individual gene trees are poorly supported. In particular, the (*Andalucia*, other jakobids, Heterolobosea, Euglenozoa) clade observed in the combined gene analysis is not present in the ML tree for any one gene, and receives, at best, very weak bootstrap support. Nonetheless, for all genes except α-tubulin, the *Andalucia* species usually branch with at least some of 'other jakobids', Heterolobosea and Euglenozoa in the ML tree, and never branch specifically with diplomonads (see below). This pattern holds irrespective of whether partial sequences with >50% of aligned sites are included or excluded. With rare exceptions, only 0–2 nodes separate *Andalucia* species from a position as the unique sister group of some or all 'other jakobids', and the intervening nodes are always very weak—most receive >30% bootstrap support (Table 1). Therefore we characterize all examined genes except α-tubulin as not being in strong conflict with the combined gene phylogeny, at least with respect to the position of *Andalucia*.

In the α-tubulin trees, by contrast, *Andalucia* branches with diplomonads, parabasalids and opisthokonts, and specifically with diplomonads. In this position, nine nodes separate *Andalucia* from any other jakobid, and three of these nodes receive >30% bootstrap support (Table 1). In particular, the (diplomonads, parabasalids, opisthokonts, *Andalucia*) grouping has a long 'basal' branch and receives 87% bootstrap support while the (diplomonads, *Andalucia*) grouping receives 59% support. Thus, uniquely among the individual gene trees examined in this study, the α-tubulin tree seems to be in marked conflict with the combined gene phylogeny with respect to the placement of *Andalucia*.

3.4. Detailed analyses of α-tubulin

We ran a series of analyses on the α-tubulin data alone to determine whether the position of *Andalucia* was robust

Table 1
Single gene analyses: comparison of the number and strength of nodes separating *Andalucia* from other jakobids

Dataset	# Taxa	# Nodes separating *Andalucia* and other jakobids	Highest bootstrap support for separating node %
Actin, all	47	0	NA
Actin, no p[a]	46	12	<30
β-tubulin, all	47	1/1[b]	<30
β-tubulin, no p[a]	46	2/2[b]	<30
EF-1α	40	1/1[b,c,d]	<30
EF-2, all	45	0/3[b]	36
EF-2, no p[a]	42	1	<30
HSP70	43	1	30
HSP90, all	47	1	<30
HSP90, no p[a]	46	1	<30
α-Tubulin	47	9[c]	87

[a] Excluding 'partial' sequences containing <50% of analysed sites.
[b] *Andalucia* not monophyletic in ML tree; node counts calculated by assuming *Andalucia* is a clade in the positions of each species.
[c] 'Other jakobids' not monophyletic in ML tree. Node count calculated for the clade of other jakobids closest to *Andalucia*.
[d] Node 'separating' *Andalucia* and (most) other jakobids is due to *Cyanophora* (a glaucophyte) falling cladistically within *Andalucia*. The *Andalucia* + *Cyanophora* clade is actually sister to the main 'other jakobid' clade in the ML tree.

to changes in taxon sampling. Firstly we analysed a larger 75-taxon dataset of α-tubulin sequences that better represented broad-scale eukaryotic diversity, but excluded long-branch-forming sequences (e.g., those from non-flagellated fungi). This dataset includes *Apusomonas*, which branches close to opisthokonts in α-tubulin trees (Kim et al., 2006), and the diplomonad relative *Carpediemonas* (see Simpson et al., 2002). Then, from both the 75-taxon dataset and the original 47-taxon dataset, we excluded either diplomonads (plus *Carpediemonas*) and parabasalids, or opisthokonts (with or without apusomonads for the 75-taxon dataset), and re-estimated the ML trees and bootstrap support values. In all cases a clade containing *Andalucia* and all included diplomonads, *Carpediemonas*, parabasalids, opisthokonts and *Apusomonas* is still recovered (Fig. 2). In fact, this clade is even stronger than in the original analysis of the 47-taxon dataset (bootstrap support 88–99%). Within this clade, the ML trees always show *Andalucia* as related to diplomonads rather than parabasalids or opisthokonts (or *Apusomonas*), although bootstrap support is never very high (44–72%). Note that when *Carpediemonas* is included, it always branches as the sister group to diplomonads (bootstrap support 85–86%), thus in these analyses *Andalucia* is sister to the diplomonad–*Carpediemonas* clade, rather than to diplomonads specifically (Fig. 2).

We observed nothing intrinsic about the *Andalucia* α-tubulin sequences to suggest that the position of *Andalucia* in α-tubulin trees might be a phylogenetic analysis artefact. For both the 47-taxon and 75-taxon datasets, all sequences within the (diplomonads, parabasalids, opisthokonts, *Andalucia*) subtree pass naïve χ^2 tests of amino acid composition, whether or not constant sites are considered. While *Andalucia*'s closest relatives in the α-tubulin tree—diplomonads—include relatively long branches, the branches of the *Andalucia* lineage are not extraordinarily long. The terminal branches of *Andalucia* sequences are actually shorter than the average among terminal branches for both the 47-taxon and 75-taxon α-tubulin datasets. The basal branch for *Andalucia* is only slightly longer (0.65 standard deviations—SD) than the median of all branches in the 47-taxon analysis (0.29 SD longer than the mean). This branch is 2.7 standard deviations longer than the median (2.4 SD from the mean) in the 75-taxon dataset, from which many long-branching sequences were excluded *a priori*, but is still not long in an absolute sense (0.083 expected substitutions per site under the WAG+Γ model), and is shorter than the basal branch for parabasalids. Thus, if long-branch attraction were structuring this part of the α-tubulin tree, it would not be expected that diplomonads (and *Carpediemonas*) would attract *Andalucia* rather than, for example, parabasalids, especially since parabasalids are probably the true closest sampled relatives of the (diplomonads, *Carpediemonas*) clade (see Simpson et al., 2006), and long-branch attraction and historical signal might be expected to reinforce each other.

Visual inspection of the α-tubulin alignment shows that *Andalucia* shares a number of unusual amino acid characters with either (i) diplomonads and parabasalids, (ii) diplomonads, parabasalids and opisthokonts, or (iii) diplomonads, parabasalids, opisthokonts and Amoebozoa [Amoebozoa appear to be closely related to opistokonts—see Discussion, and some Amoebozoa, e.g., *Mastigamoeba*, tend to branch close to the (opisthokont, diplomonad, parabasalid, *Andalucia*) grouping in α-tubulin trees—see Fig. 2]. Curiously, many of these characters are located in one 15-residue segment of the sequence. Nonetheless, when this region is excluded from analysis, the (opisthokont, diplomonad, parabasalid, *Andalucia*) clade is still recovered. Bootstrap support for this clade remains strong in the 75-taxon dataset (96%). Support is lower with the 47-taxon dataset (63%), however this is due to variability in the positions of the long-branching sequences from the fungi *Neurospora* and *Cryptococcus*, not the *Andalucia* sequences—if *Neurospora* and *Cryptococcus* are simply pruned from the bootstrap trees, the (opisthokont, diplomonad, parabasalid, *Andalucia*) clade receives strong support (91%). In both analyses *Andalucia* still branches with diplomonads (and *Carpediemonas*), with moderate support (71% for the 75-taxon dataset, 61% with the 47-taxon dataset).

3.5. Andalucia monophyly in single gene analyses

One other noteworthy difference between α-tubulin and the other single gene datasets concerns the support for the monophyly of *Andalucia*. *Andalucia* is not recovered as a clade in the ML trees for β-tubulin, EF-1α and EF-2, and bootstrap support for *Andalucia* monophyly is very

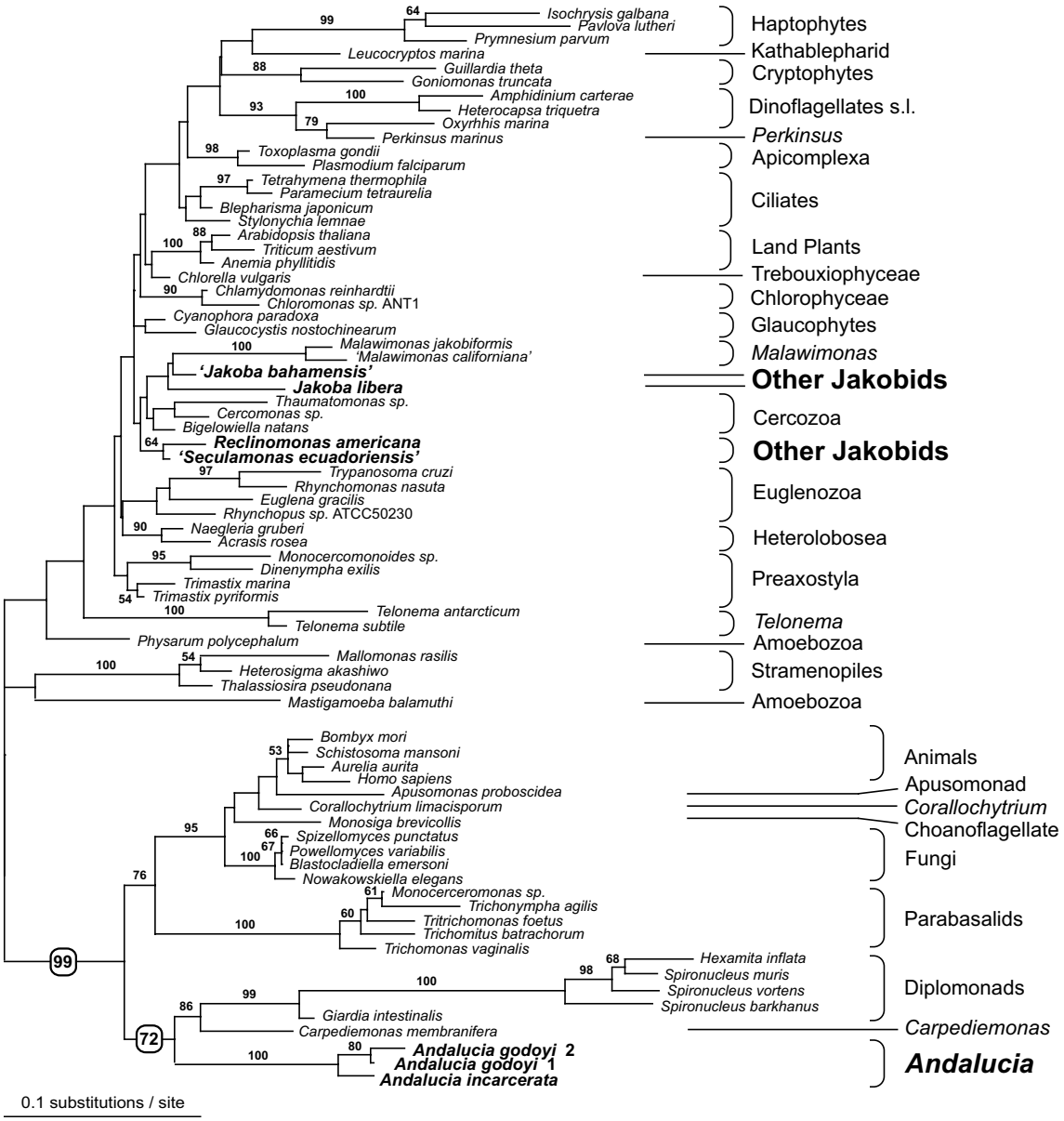

Fig. 2. Maximum likelihood tree for the 75-taxon α-tubulin dataset (RAxML, WAG+Γ model), with percentage bootstrap support values depicted (PHYML, WAG+Γ model, 1000 replicates). Bootstrap values <50% are not shown. Note the position of *Andalucia*—closely related to diplomonads, within a strongly supported clade also containing parabasalids and opisthokonts.

low (13–36%). *Andalucia* is monophyletic in the ML trees for actin, HSP90 and α-tubulin. Bootstrap support for this clade is weak (42% or 43%) with actin, and only moderately strong (71% or 77%) with HSP90, whereas it is very strong (99–100%) in all of the analyses of α-tubulin data described above. In the cases of actin and HSP90, the ML tree for the gene resembles the ML tree for the combined protein datasets in showing a deep divergence between the two *Andalucia* species, with the branch at the base of the clade being much shorter than the terminal branches (data not shown). By contrast, the *Andalucia* clade is quite shallow in α-tubulin trees, with the basal branch for the clade being several times longer than the terminal branches (see Fig. 2).

3.6. Combined gene phylogeny excluding α-tubulin

If α-tubulin is excluded from the original 47-taxon alignment, the ML trees are similar in most respects to those for the full seven-protein analyses (Fig. 3). The one major exception is that diplomonads and parabasalids are no longer close to opisthokonts but instead fall as the sister group to Preaxostyla with considerable bootstrap support (96% PHYML; 78% RAxML). As a result, opisthokonts and Amoebozoa now form a strongly supported clade (92% PHYML; 91% RAxML). *Malawimonas* branches with (*Andalucia*, other jakobids, Heterolobosea, Euglenozoa) in both the concatenated and unlinked ML trees, but support is weak. The (*Andalucia*, other jakobids,

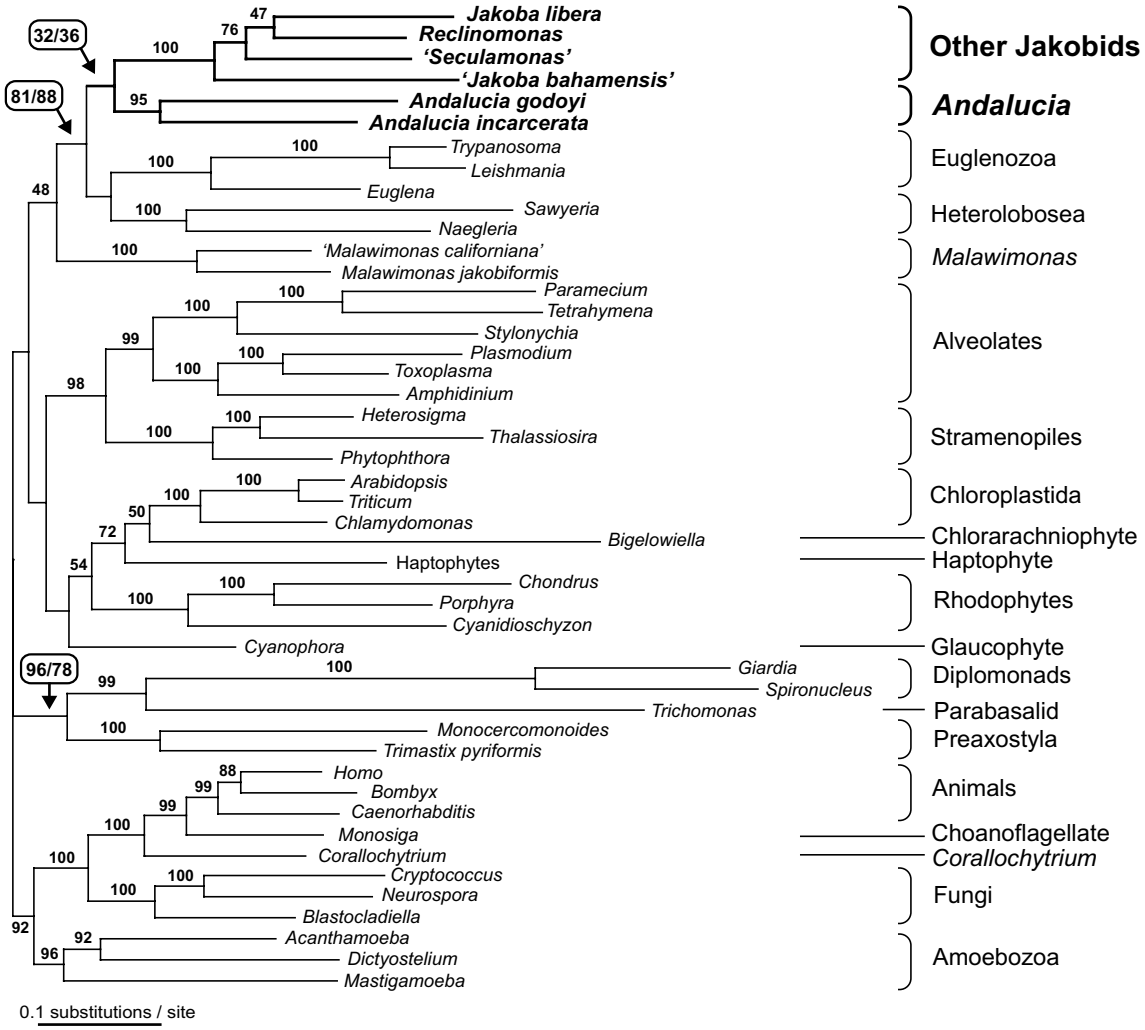

Fig. 3. Maximum likelihood tree for the six-protein dataset (i.e., with α-tubulin excluded) under the concatenated ('super-gene') model (WAG+Γ). Percentage bootstrap support values are depicted. Single numbers represent values estimated using PHYML (WAG+Γ, 500 replicates). Most values <40% are not shown. For important nodes, double numbers are shown giving the PHYML values followed by values estimated using RAxML (WAG+CAT, 200 replicates). There were no important differences between the ML trees for the concatenated and unlinked models.

Heterolobosea, Euglenozoa) clade itself now receives consistently strong bootstrap support (81% PHYML; 88% RAxML). However, while *Andalucia* again branches as the sister to other jakobids, bootstrap support for this clade remains weak (32% PHYML; 36% RAxML).

With α-tubulin excluded, the ML tree for the 54-taxon dataset is completely consistent with the concatenated model tree for the 47-taxon dataset. Bootstrap support for the (*Andalucia*, other jakobids, Heterolobosea, Euglenozoa) clade is moderately strong (74%), and support for (*Andalucia*, other jakobids) is again low (36%). The support for *Andalucia* monophyly is 94%.

4. Discussion

4.1. The phylogenetic position of Andalucia

Previous studies based on one or two genes have not resolved the phylogenetic position of *Andalucia* (Edgcomb et al., 2001; Lara et al., 2006; Simpson et al., 2002). The current study somewhat improves our understanding—our multi-protein datasets provide the first substantial support for a grouping of *Andalucia* with other jakobids, Euglenozoa and Heterolobosea, especially once the aberrant signal from α-tubulin is excluded. Also for the first time, our ML trees place *Andalucia* specifically with other jakobids, as suggested by morphological comparisons. Nonetheless, bootstrap support for this larger jakobid clade is negligible in all our analyses. Even if *Andalucia* is specifically related to other jakobids, it appears that their divergence occurred deep in the history of this lineage, and relatively little signal remains in individual gene sequences. Despite the very strong morphological similarity between *Andalucia* and other jakobids, conclusive resolution of their precise relationship could well require sequence data from dozens of genes.

4.2. The evolutionary history of α-tubulin

The α-tubulin dataset clearly includes a very different phylogenetic signal from the other protein sequences examined,

placing *Andalucia* together with opisthokonts, diplomonads and parabasalids, rather than close to other jakobids. This signal is robust to changes in taxon sampling, including the exclusion of long-branching immediate relatives of *Andalucia* α-tubulin (diplomonads and parabasalids). The *Andalucia* α-tubulins are not especially 'long branches', nor do they display unusual amino acid compositions. The signal is not isolated to one segment of the sequence (or at least, to the most conspicuous 'potentially signal-rich' segment), and therefore is unlikely to result from partial gene conversion. Thus, there is strong evidence that *Andalucia* α-tubulin genes have a genuinely different history from that of the other genes from *Andalucia* examined here.

The simplest explanation is that a common ancestor of *Andalucia godoyi* and *Andalucia incarcerata* was the recipient in a eukaryote-to-eukaryote lateral transfer of an α-tubulin gene. The donor in this transfer would appear to have been an organism from the opisthokont lineage, or from the diplomonad–parabasalid clade. In light of the persistent (though never strong) signal uniting *Andalucia* and diplomonads, the donor was most likely an early branch from the diplomonad lineage.

In most single gene trees and concatenated gene trees in which *Andalucia* is monophyletic, the group is a 'deep' clade with a relatively short basal branch (this study; Lara et al., 2006, 2007). In α-tubulin trees, by contrast, *Andalucia* forms a shallow clade, with a proportionately long basal branch. Assuming that the *Andalucia* α-tubulin genes were acquired by LGT, there are two interesting explanations possible for this relatively long basal branch. One possibility is that there was a high rate of evolution in *Andalucia* α-tubulin genes immediately after transfer, perhaps associated with adaptation to a novel cellular milieu. A second possibility stems from the fact that the branch leading to *Andalucia* represents a combination of the lineage of *Andalucia* from the LGT event through to the divergence of the lineages of the two sampled species, *plus* the lineage of the donor, back to the last common ancestor that it shares with any organism that was included in our α-tubulin analysis (excluding gene duplications, lineage sorting, etc.). If the donor lineage had been independent of the other sampled lineages for a long time and/or was rapidly evolving, much of the basal branch of *Andalucia* in α-tubulin trees could actually represent the period that this gene was resident in the genome of the donor lineage, and not the genome of direct ancestors of *Andalucia*. Testing these possibilities is complicated by several factors. Firstly, the most likely close relatives of the donor lineage—diplomonads (and *Carpediemonas*), or perhaps parabasalids—are themselves rapidly evolving. Secondly, it is possible that diplomonads and parabasalids also acquired their α-tubulin genes through an earlier LGT (see below). Thirdly, our taxon sampling at the base of the diplomonad–*Carpediemonas* clade is poor—it would be interesting to compare the α-tubulin genes of *Andalucia* to those of other deep-branching relatives of diplomonads, such as the recently described *Dysnectes* (Yubuki et al., 2007).

In addition to the substantial case that *Andalucia* acquired α-tubulin genes by lateral transfer, this study highlights the possibility that diplomonads and parabasalids also acquired their α-tubulin genes by this mechanism. Previous studies have noted that the strong placement of diplomonads and parabasalids with opisthokonts in α-tubulin trees is discordant with the phylogenetic signal from other genes (Arisue et al., 2005; Hampl et al., 2005; Simpson et al., 2006), however, explaining this disparity has been complicated by uncertainty over the phylogenetic position of diplomonads and parabasalids. Some studies have recovered a specific relationship between diplomonads, parabasalids and Preaxostyla, but support has been weak (Cavalier-Smith, 2003; Simpson et al., 2002, 2006), or taxon sampling has been low (Hampl et al., 2005). Despite our dataset being quite similar to that of Simpson et al. (2006), we unexpectedly found strong support for a (diplomonads, parabasalids, Preaxostyla) clade once α-tubulin data is excluded. This strengthens the hypothesis that α-tubulin trees are not reflecting organismal phylogeny in their placement of diplomonads and parabasalids relative to opisthokonts. Given the lack of obvious reasons to suspect phylogenetic analysis artefact (for example, the relationship remains strong even when long-branching opisthokonts are excluded), LGT is perhaps the most likely explanation for this disparity.

Although the evidence is very weak, it seems more likely that diplomonads and parabasalids were the recipients of the putative LGT, and a relative of opisthokonts was the donor, rather than the other way around. Multi-gene phylogenies and other molecular data indicate that opisthokonts and Amoebozoa are closely related (e.g., Baldauf et al., 2000; Bapteste et al., 2002; Richards and Cavalier-Smith, 2005; Rodríguez-Ezpeleta et al., 2007). Shorter-branching Amoebozoa, such as *Mastigamoeba* and *Physarum* tend to branch close to opisthokonts, diplomonads and parabasalids in α-tubulin trees, albeit with weak support, which is consistent with opisthokonts being related to the donor lineage. On the other hand, parabasalids are closer to opisthokonts than to diplomonads in our α-tubulin trees, which is consistent with LGT from a donor within the diplomonad–parabasalid clade to an ancestor of opisthokonts. This topology, however, could also be explained by separate transfers into ancestors of diplomonads and parabasalids, or simply by incorrect phylogenetic reconstruction. It is noted that transfers of other genes have been traced both to a common ancestor of diplomonads and parabasalids (Andersson et al., 2005; Henze et al., 2001) and to the opisthokont stem lineage (Huang et al., 2005).

Yet another LGT involving α-tubulin is hinted at by this study. In our α-tubulin trees the apusomonad *Apusomonas* falls within the (opisthokonts, diplomonads, parabasalids) clade with strong support, and branches close to, or actually within opisthokonts. The phylogenetic position of apusomonads is uncertain, but it has been noted previously that α-tubulin is the only gene (of the tiny number studied) that by itself strongly places *Apusomonas* with

opisthokonts (Kim et al., 2006). It is difficult at present to test the possibility of LGT from opisthokonts to *Apusomonas*, since the best current hypothesis is that apusomonads are closely related to opisthokonts (see Kim et al., 2006). α-Tubulin data from other apusomonads may help in exploring this question.

5. Conclusions

The α-tubulin of *Andalucia* may represent only the most clear of two or more instances of lateral transfer of characteristic 'opisthokont-type' α-tubulin genes between eukaryotic lineages. Although less extreme, the case of α-tubulin bears similarities to that of EF-1α/EFL, where numerous distantly related lineages have independently 'replaced' an 'essential' gene (EF-1α) with a distinctly different, though related gene (EFL) that was apparently acquired by LGT following its evolution in one eukaryotic lineage (Keeling and Inagaki, 2004). In some groups there is evidence for the ancestral presence of both EF-1α and EFL, followed by differential loss in descendent lineages (Noble et al., 2007), a process for which we have no evidence in the case of α-tubulin. As with EF-1α/EFL, it would be interesting to know whether opisthokont-type α-tubulin conferred a selective advantage on its recipients. For the moment, however, we emphasise that α-tubulin has undergone LGT between distantly related microbial eukaryotes with no known history of eukaryote–eukaryote endosymbiosis. This propensity should be taken into account when inferring deep-level eukaryote evolution using this marker—in particular, studies employing a 'few genes, many species' approach could be significantly affected.

Acknowledgments

The research was funded by NSERC Grant 298366-04 to AGBS. Computational resources were funded by Genome Atlantic. AGBS thanks the Canadian Institute for Advanced Research (CIFAR), programs in Evolutionary Biology and Integrated Microbial Biodiversity, for support as a 'Scholar'. Special thanks to Andrew Roger and Vladimir Hampl (Dalhousie University) for unpublished sequences from *Mastigamoeba* and *Monocercomonoides,* and to B. Franz Lang (Université de Montréal) for jakobid and *Malawimonas* data. Thanks to Andrew Roger for discussions and Nicole White for proof-reading.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.ympev.2007.11.035.

References

Adl, S.M., Simpson, A.G.B., Farmer, M.A., Andersen, R.A., Anderson, O.R., Barta, J.R., Bowser, S.S., Brugerolle, G., Fensome, R.A., Fredericq, S., James, T.Y., Karpov, S., Kugrens, P., Krug, J., Lane, C.E., Lewis, L.A., Lodge, J., Lynn, D.H., Mann, D.G., McCourt, R.M., Mendoza, L., Moestrup, O., Mozley-Standridge, S.E., Nerad, T.A., Shearer, C.A., Smirnov, A.V., Spiegel, F.W., Taylor, M.F.J.R., 2005. The new higher-level classification of eukaryotes with emphasis on the taxonomy of protists. J. Eukaryot. Microbiol. 52, 399–451.

Andersson, J.O., 2005. Lateral gene transfer in eukaryotes. Cell. Mol. Life Sci. 11, 1182–1197.

Andersson, J.O., Roger, A.J., Sarchfield, S.W., 2005. Gene transfers from Nanoarchaeota to an ancestor of diplomonads and parabasalids. Mol. Biol. Evol. 22, 85–90.

Arisue, N., Hasegawa, M., Hashimoto, T., 2005. Root of the Eukaryota tree as inferred from combined maximum likelihood analyses of multiple molecular sequence data. Mol. Biol. Evol. 22, 409–420.

Baldauf, S.L., 2003. The deep roots of eukaryotes. Science 300, 1703–1706.

Baldauf, S.L., Palmer, J.D., 1993. Animals and Fungi are each others closest relatives—congruent evidence from multiple proteins. Proc. Natl. Acad. Sci. USA 90, 11158–11562.

Baldauf, S.L., Roger, A.J., Wenk-Siefert, I., Doolittle, W.F., 2000. A kingdom-level phylogeny of eukaryotes based on combined protein data. Science 290, 972–977.

Bapteste, E., Brinkmann, H., Lee, J.A., Moore, D.V., Sensen, C.W., Gordon, P., Durufle, L., Gaasterland, T., Lopez, P., Müller, M., Philippe, H., 2002. The analysis of 100 genes supports the grouping of three highly divergent amoebae: *Dictyostelium*, *Entamoeba*, and *Mastigamoeba*. Proc. Natl. Acad. Sci. USA 99, 1414–1419.

Behnke, A., Bunge, J., Barger, K., Breiner, H.W., Alla, V., Stoeck, T., 2006. Microeukaryote community patterns along an O_2/H_2S gradient in a supersulfidic anoxic fjord (Framvaren, Norway). Appl. Environ. Microbiol. 72, 3626–3636.

Bernard, C., Simpson, A.G.B., Patterson, D.J., 2000. Some free-living flagellates (Protista) from anoxic habitats. Ophelia 52, 113–142.

Berney, C., Fahrni, J., Pawlowski, J., 2004. How many novel eukaryotic "kingdoms"? Pitfalls and limitations of environmental DNA surveys. BMC Biol. 2, 13.

Burki, F., Pawlowski, J., 2006. Monophyly of Rhizaria and multigene phylogeny of unicellular bikonts. Mol. Biol. Evol. 23, 1922–1930.

Burki, F., Shalchian-Tabrizi, K., Minge, M., Skjaeveland, A., Nikolaev, S.I., Jakobsen, K.S., Pawlowski, J., 2007. Phylogenomics reshuffles the eukaryotic supergroups. PLoS ONE 2, 790.

Cavalier-Smith, T., 2003. The excavate protozoan phyla Metamonada Grasse emend. (Anaeromonadea, Parabasalia, *Carpediemonas*, Eopharyngia) and Loukozoa emend. (Jakobea, *Malawimonas*): their evolutionary affinities and new higher taxa. Int. J. Syst. Evol. Microbiol. 53, 1741–1758.

Cavalier-Smith, T., 2004. Only six kingdoms of life. Proc. Royal Soc. Lond. B 271, 1251–1262.

Ciccarelli, F.D., Doerks, T., von Mering, C., Creevey, C.J., Snel, B., Bork, P., 2006. Toward automatic reconstruction of a highly resolved tree of life. Science 311, 1283–1287.

Clark, C.G., Diamond, L.S., 1991. Ribosomal RNA genes of pathogenic and non-pathogenic *Entamoeba histolytica* are distinct. Mol. Biochem. Parasitol. 49, 297–302.

Edgcomb, V.P., Roger, A.J., Simpson, A.G.B., Kysela, D., Sogin, M.L., 2001. Evolutionary relationships among "jakobid" flagellates as indicated by alpha- and beta-tubulin phylogenies. Mol. Biol. Evol. 18, 514–522.

Edlind, T.D., Li, J., Visvesvara, G.S., Vodkin, M.H., McLaughlin, G.L., Katiyar, S.K., 1996. Phylogenetic analysis of β-tubulin sequences from amitochondrial protozoa. Mol. Phylogenet. Evol. 5, 359–367.

Gile, G.H., Patron, N.J., Keeling, P.J., 2006. EFL GTPase in cryptomonads and the distribution of EFL and EF-1α in chromalveolates. Protist 157, 435–444.

Guindon, S., Gascuel, O., 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst. Biol. 52, 696–704.

Hackett, J.D., Yoon, H.S., Li, S., Reyes-Prieto, A., Rümmele, S.E., Bhattacharya, D., 2007. Phylogenomic analysis supports the mono-

phyly of cryptophytes and haptophytes and the association of 'Rhizaria' with chromalveolates. Mol. Biol. Evol. 24, 1702–1713.
Hampl, V., Horner, D.S., Dyal, P., Kulda, J., Flegr, J., Foster, P., Embley, T.M., 2005. Inference of the phylogenetic position of oxymonads based on nine genes: support for Metamonada and Excavata. Mol. Biol. Evol. 22, 2508–2518.
Harper, J.T., Waanders, E., Keeling, P.J., 2005. On the monophyly of chromalveolates using a six-protein phylogeny of eukaryotes. Int. J. Syst. Evol. Microbiol. 55, 487–496.
Henze, K., Horner, D.S., Suguri, S., Moore, D.V., Sanchez, L.B., Müller, M., Embley, T.M., 2001. Unique phylogenetic relationships of glucokinase and glucosephosphate isomerase of the amitochondriate eukaryotes *Giardia intestinalis*, *Spironucleus barkhanus* and *Trichomonas vaginalis*. Gene 281, 123–131.
Huang, J., Xu, Y., Gogarten, J.P., 2005. The presence of a haloarchaeal type tyrosyl-tRNA synthetase marks the opisthokonts as monophyletic. Mol. Biol. Evol. 22, 2142–2146.
Jobb, G., von Haeseler, A., Strimmer, K., 2004. TREEFINDER: a powerful graphical analysis environment for molecular phylogenetics. BMC Evol. Biol. 4, 18.
Keeling, P.J., 2001. Foraminifera and Cercozoa are related in actin phylogeny: two orphans find a home? Mol. Biol. Evol. 18, 1551–1557.
Keeling, P.J., Burger, G., Durnford, D.G., Lang, B.F., Lee, R.W., Pearlman, R.E., Roger, A.J., Gray, M.W., 2005. The tree of eukaryotes. Trends Ecol. Evol. 20, 670–676.
Keeling, P.J., Doolittle, W.F., 1996. Alpha-tubulin from early-diverging eukaryotic lineages and the evolution of the tubulin family. Mol. Biol. Evol. 13, 1297–1305.
Keeling, P.J., Inagaki, Y., 2004. A class of eukaryotic GTPase with a punctate distribution suggesting multiple functional replacements of translation elongation factor 1 alpha. Proc. Natl. Acad. Sci. USA 101, 15380–15385.
Kim, E., Simpson, A.G.B., Graham, L.E., 2006. Evolutionary relationships of apusomonads inferred from taxon-rich analyses of six nuclear-encoded genes. Mol. Biol. Evol. 23, 2455–2466.
Lara, E., Berney, C., Ekelund, F., Harms, H., Chatzinotas, 1 A., 2007. Molecular comparison of cultivable protozoa from a pristine and a polycyclic aromatic hydrocarbon polluted site. Soil. Biol. Biochem. 39, 139–148.
Lara, E., Chatzinotas, A., Simpson, A.G.B., 2006. *Andalucia* (gen. nov,): a new taxon for the deepest branch within jakobids (Jakobida; Excavata), based on morphological and molecular study of a new flagellate from soil. J. Eukaryot. Microbiol. 53, 112–120.
Li, J., Katiyar, S.K., Hamelin, A., Visvesvara, G.S., Edlind, T.D., 1996. Tubulin genes from AIDS-associated microsporidia and implications for phylogeny and benzimidazole sensitivity. Mol. Biochem. Parasitol. 78, 289–295.
Luo, Q.W., Krumholz, L.R., Najar, F.Z., Peacock, A.D., Roe, B.A., White, D.C., Elshahed, M.S., 2005. Diversity of the microeukaryotic community in sulfide-rich Zodletone Spring (Oklahoma). Appl. Environ. Microbiol. 71, 6175–6184.
Medlin, L., Elwood, H.J., Stickel, S., Sogin, M.L., 1988. The characterization of enzymatically amplified eukaryotic 16S-like rRNA-coding regions. Gene 71, 491–499.
Nikolaev, S.I., Berney, C., Fahrni, J.F., Bolivar, I., Polet, S., Mylnikov, A.P., Aleshin, V.V., Petrov, N.B., Pawlowski, J., 2004a. The twilight of Heliozoa and rise of Rhizaria, an emerging supergroup of amoeboid eukaryotes. Proc. Natl. Acad. Sci. USA 101, 8066–8071.
Nikolaev, S.I., Mylnikov, A.P., Berney, C., Fahrni, J., Pawlowski, J., Aleshin, V.V., Petrov, N.B., 2004b. Molecular phylogenetic analysis places *Percolomonas cosmopolitus* within Heterolobosea: evolutionary implications. J. Eukaryot. Microbiol. 51, 575–581.
Noble, G.P., Rogers, M.B., Keeling, P.J., 2007. Complex distribution of EFL and EF-1alpha proteins in the green algal lineage. BMC Evol. Biol. 7, 82.
Nozaki, H., Iseki, M., Hasegawa, M., Misawa, K., Nakada, T., Sasaki, N., Watanabe, M., 2007. Phylogeny of primary photosynthetic eukaryotes as deduced from slowly evolving nuclear genes. Mol. Biol. Evol. 24, 1592–1595.
Parfrey, L.W., Barbero, E., Lasser, E., Dunthorn, M., Bhattacharya, D., Patterson, D.J., Katz, L.A., 2006. Evaluating support for the current classification of eukaryotic diversity. PLoS Genet. 2, 2062–2073.
Patron, N.J., Inagaki, Y., Keeling, P.J., 2007. Multiple gene phylogenies support the monophyly of cryptomonad and haptophyte host lineages. Curr. Biol. 17, 887–891.
Philippe, H., Snell, E.A., Bapteste, E., Lopez, P., Holland, P.W.H., Casane, D., 2004. Phylogenomics of eukaryotes: impact of missing data on large alignments. Mol. Biol. Evol. 21, 1740–1752.
Richards, T.A., Cavalier-Smith, T., 2005. Myosin domain evolution and the primary divergence of eukaryotes. Nature 436, 1113–1118.
Rodríguez-Ezpeleta, N., Brinkmann, H., Burey, S.C., Roure, B., Burger, G., Loffelhardt, W., Bohnert, H.J., Philippe, H., Lang, B.F., 2005. Monophyly of primary photosynthetic eukaryotes: green plants, red algae, and glaucophytes. Curr. Biol. 15, 1325–1330.
Rodríguez-Ezpeleta, N., Brinkmann, H., Burger, G., Roger, A.J., Gray, M.W., Philippe, H., Lang, B.F., 2007. Toward resolving the eukaryotic tree: the phylogenetic positions of jakobids and cercozoans. Curr. Biol. 17, 1420–1425.
Ronquist, F., Huelsenbeck, J.P., 2003. MRBAYES 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19, 1572–1574.
Shalchian-Tabrizi, K., Eikrem, W., Klaveness, D., Vaulot, D., Minge, M.A., Le Gall, G., Romari, K., Throndsen, J., Botnen, A., Massana, T., Thomsen, H.A., Jakobsen, K.S., 2006. Telonemia, a new protist phylum with affinity to chromist lineages. Proc. R. Soc. Lond. B Biol. Sci. 273, 1833–1842.
Simpson, A.G.B., 2003. Cytoskeletal organisation phylogenetic affinities and systematics in the contentious taxon Excavata (Eukaryota). Int. J. Syst. Evol. Microbiol. 53, 1759–1777.
Simpson, A.G.B., Inagaki, Y., Roger, A.J., 2006. Comprehensive multigene phylogenies of excavate protists reveal the evolutionary positions of "primitive" eukaryotes. Mol. Biol. Evol. 23, 615–625.
Simpson, A.G.B., Patterson, D.J., 2001. On core jakobids and excavate taxa: the ultrastructure of *Jakoba incarcerata*. J. Eukaryot. Microbiol. 48, 480–492.
Simpson, A.G.B., Roger, A.J., 2004. The real 'kingdoms' of eukaryotes. Curr. Biol. 14, R693–R696.
Simpson, A.G.B., Roger, A.J., Silberman, J.D., Leipe, D.D., Edgcomb, V.P., Jermiin, L.S., Patterson, D.J., Sogin, M.L., 2002. Evolutionary history of "early diverging" eukaryotes: the excavate taxon *Carpediemonas* is a close relative of *Giardia*. Mol. Biol. Evol. 19, 1782–1791.
Stamatakis, A., 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22, 2688–2690.
Steenkamp, E.T., Wright, J., Baldauf, S.L., 2006. The protistan origins of animals and fungi. Mol. Biol. Evol. 23, 93–106.
Strimmer, K., von Haeseler, A., 1996. Quartet puzzling: a quartet maximum-likelihood method for reconstructing tree topologies. Mol. Biol. Evol. 13, 964–969.
Yubuki, N., Inagaki, Y., Nakayama, T., Inouye, I., 2007. Ultrastructure and ribosomal RNA phylogeny of the free-living heterotrophic flagellate *Dysnectes brevis* n. gen., n. sp., a new member of the Fornicata. J. Eukaryot. Microbiol. 54, 191–200.