Ancient Origin of Glycosyl Hydrolase Family 9 Cellulase Genes

Angus Davison† and Mark Blaxter**

*Institute of Evolutionary Biology, School of Biological Sciences, University of Edinburgh, Edinburgh, United Kingdom; and †Institute of Genetics, University of Nottingham, Nottingham, United Kingdom

While it is widely accepted that most animals (Metazoa) do not have endogenous cellulases, relying instead on intestinal symbionts for cellulose digestion, the glycosyl hydrolase family 9 (GHF9) cellulases found in the genomes of termites, abalone, and sea squirts could be an exception. Using information from expressed sequence tags, we show that GHF9 genes (subgroup E2) are widespread in Metazoa because at least 11 classes in five phyla have expressed GHF9 cellulases. We also demonstrate that eukaryotic GHF9 gene families are ancient, forming distinct monophyletic groups in plants and animals. As several intron positions are also conserved between four metazoan phyla then, contrary to the still widespread belief that cellulases were horizontally transferred to animals relatively recently, GHF9 genes must derive from an ancient ancestor. We also found that sequences isolated from the same animal phylum tend to group together, and in some deuterostomes, GHF9 genes are characterized by substitutions in catalytically important sites. Several paralogous subfamilies of GHF9 can be identified in plants, and genes from primitive species tend to arise basally to angiosperm representatives. In contrast, GHF9 subgroup E2 genes are relatively rare in bacteria.

Introduction

Cellulose is the most abundant organic compound on Earth. Therefore, to understand global carbon cycling the dynamics of cellulose synthesis and degradation must be understood. Plants, some bacteria, fungi, protozoa, and sea squirts (ascidians) synthesize cellulose and also need to be able to degrade or modify it during growth and development. The majority of decomposing degradation is carried out by bacteria, fungi, and protozoa, most famously as commensals in the guts of herbivorous animals. In consequence, it is commonly believed (e.g., Morris 2003) that most animals are unable to digest cellulose except when assisted by these commensals and that ''surprising'' exceptions in termites, nematodes, and sea squirts have acquired their cellulolytic endoglucanases by horizontal gene transfer from prokaryotes (Smant et al. 1998; Watanabe et al. 1998; Dehal et al. 2002; Pennisi 2002; Scholl et al. 2003). The alternative explanation for the presence of cellulases in these diverse animals is that they are derived from genes in an ancient ancestral eukaryote and have persisted only in some metazoan lineages (Yan et al. 1998; Lo, Watanabe, and Sugimura 2003; Matthysse et al. 2004; Nakashima et al. 2004).

Before concluding that genes have been gained by horizontal transfer, it is necessary to rigorously investigate the evidence, preferably using a gene-by-gene approach (Ochman, Lawrence, and Groisman 2000; Genereux and Logsdon 2003). Fourteen families of glycosyl hydrolases (GHF) are able to degrade cellulose (GHF5, 6, 7, 8, 9, 10, 12, 26, 44, 45, 48, 51, 61, and 74; Henrissat 1991; see http://afmb.cnrs-mrs.fr/CAZY/index.html). Five of these families have representatives in Metazoa (table 1). For four (GHF5, GHF6, GHF10, GHF45), very few animal-derived members have been identified. Tylenchine plant-parasitic nematodes (Smant et al. 1998) and a phytophagous beetle (Sugimura et al. 2003) express GHF5 cellulases (table 1). There is reasonable phylogenetic evidence that both of these genes are derived from bacteria by horizontal gene transfer (Yan et al. 1998; Lo, Watanabe, and Sugimura 2003). The sea squirts *Ciona intestinalis* and *Ciona savignyi* have a protein with a putative GHF6-like domain (Matthysse et al. 2004; Nakashima et al. 2004). Again, there is reasonable phylogenetic evidence that the GHF6-like domain was gained by horizontal transfer (Matthysse et al. 2004; Nakashima et al. 2004). Finally, GHF45 cellulases have been described from a beetle (Girard and Jouanin 1999) and two mollusks (Xu, Janson, and Sellos 2001; Harada, Hosoiri, and Kuroda 2004), and a GHF10 cellulase has been isolated from a mollusk (Wang et al. 2003) (table 1). Phylogenetic analysis to test for an ancient origin using these genes is compromised by a lack of data. Even in the case of GHF5 and GHF6 genes, phylogenetic resolution is quite poor, presumably because the genes are short and saturated for substitution (Lo, Watanabe, and Sugimura 2003; Matthysse et al. 2004; Nakashima et al. 2004). However, the fifth family of metazoan glycosyl hydrolase genes—GHF9 endo-beta-1,4-glucanases—is exceptional because the core gene sequence is both relatively long (over 430 amino acids) and conserved.

GHF9 has been relatively widely studied in the Metazoa, following the surprising discovery of endogenous GHF9 genes in termites (phylum Arthropoda; Watanabe et al. 1998; Watanabe and Tokuda 2001). Initially, their origin in the arthropods was attributed to a date before the divergence of termites and cockroaches, approximately 250 MYA. GHF9 genes have recently been reported in two further animal phyla, the Mollusca (Suzuki, Ojima, and Nishita 2003) and Chordata (Dehal et al. 2002). GHF9 genes also have a wide distribution in angiosperms (flowering plants) and have been discovered in some fungi (Steenbakkers et al. 2002) and a single amoebozoan (*Dictyostelium discoideum*; Libertini, Li, and McQueen-Mason 2004). There are two distantly related families of the GHF9 gene: subgroup E1 is confined to bacteria (Tomme, Warren, and Gilkes 1995), whereas subgroup E2 has been found in bacteria, *Dictyostelium*, termites and other Metazoa (Tomme, Warren, and Gilkes 1995; Tokuda et al. 1999). In plants, phylogenetic analyses of GHF9 genes (subgroup

Key words: cellulase, expressed sequence tag, glycosyl hydrolase, horizontal gene transfer.

E-mail: angus.davison@nottingham.ac.uk.

Mol. Biol. Evol. 22(5):1273–1284. 2005
doi:10.1093/molbev/msi107
Advance Access publication February 9, 2005

Table 1
Metazoan Cellulases (Except GHF9)

Family	Species		GenBank	Reference
GHF5	*Psacothea hilaris*	Phytophagous beetle	AB080266	Sugimura et al. (2003)
	Globodera rostochiensis	Plant-parasitic nematode	AF004523, AF004716	Smant et al. (1998)
	Heterodera glycines	Plant-parasitic nematode	AF006052–AF006053	Smant et al. (1998)
	Meloidogyne incognita	Root-knot nematode	AF323087	T. N. Ledger, S. Jaubert, J. Cazot, L. Arhaud, P. Abad, and M. N. Rosso (personal communication)
GHF45	*Phaedon cochleariae*	Phytophagous beetle	CAA76931	Girard and Jouanin (1999)
	Apriona germari	Phytophagous beetle	AAR22385	Lee et al. (2004)
	Ips pini[a]	Phytophagous beetle	CB408544, CB408403	Eigenheer et al. (2003)
	Mytilus edulis	Mussel	CAC59694–CAC59695	Xu, Janson, and Sellos (2001)
	Lymnaea stagnalis	Snail	AB159152	Harada, Hosoiri, and Kuroda (2004)
	Bursaphelenchus xylophilus	Plant-parasitic nematode	BAD34543–BAD34548	Kikuchi et al. (2004)
	Hypsibius dujardini[a]	Tardigrade	CD449425	J. Daub, F. Thomas, A. Aboobaker, and M. L. Blaxter (personal communication)
GHF10	*Ampullaria crossean*	Snail	AAP31839	Wang et al. (2003)
GHF6	*Ciona intestinalis*	Sea squirt	AB104509	Nakashima et al. (2004)
	Ciona savignyi	Sea squirt	AY504665	Matthysse et al. (2004)

[a] EST evidence.

E2) were used to link subfamilies to specific gene function (e.g., cellulose-assisted abscission, ripening, etc; Libertini, Li, and McQueen-Mason 2004). GHF9 phylogeny has also been examined within the termites (Tokuda et al. 2004).

Lo, Watanabe, and Sugimura (2003) presented evidence, based on a conserved intron position, that the GHF9 genes of termites, abalone, and sea squirts are derived from an ancestral gene in the last common ancestor of protostomes and deuterostomes. We reasoned that if metazoan GHF9 cellulases do have a common origin in a metazoan ancestor, then it should be possible to identify GHF9 cellulase genes in the genome data that is emerging from a wide diversity of animals and other eukaryotes and use phylogenetic analysis to demonstrate an ancient endogenous origin. We show here that GHF9 endoglucanases are indeed widespread in Eukaryota and that their phylogeny strongly suggests their presence in an ancient eukaryotic ancestor.

Materials and Methods

Extraction of Sequences from Databases

Database searching was carried out during March to September 2004. Novel GHF9 cellulases were identified in GenBank (http://www.ncbi.nlm.nih.gov) by Blast searches with a variety of seed sequences previously identified as GHF9 genes. Representative sequences from all previously characterized GHF9 (subgroup E2) cellulases in bacteria, plants, and fungi were downloaded from the CAZY glycosyl hydrolase database (http://afmb.cnrs-mrs.fr/~cazy/CAZY/index.html). Several putative cellulases were also identified by searching unassembled genome sequences held on organism-specific web pages and unfinished high-throughput genome sequences. To achieve this, the following websites were used: Joint Genome Institute (http://www.jgi.doe.gov/index.html), the Wellcome Trust Sanger Institute (http://www.sanger.ac.uk), the Institute for Genome Research (http://www.tigr.org/tdb/), Washington University Genome Sequencing Centre (http://www.genome.wustl.edu), H-invitational database (http://h-invitational.jp), Baylor College of Medicine (http://www.hgsc.bcm.tmc.edu), Dictybase (http://dictybase.org), *Ciona intestinalis* genome (http://genome.jgi-psf.org/ciona/), *Apis mellifera* genome (http://hgsc.bcm.tmc.edu/projects/honeybee), and Lumbribase (http://www.earthworms.org).

Sequence Alignment and Phylogenetic Analysis

As horizontal transfer is a rare event compared with vertical transfer, even in bacteria, any given pair of genes is considerably more likely to be related by vertical rather than horizontal descent. We therefore consider vertical descent to be the null hypothesis against which alternate hypotheses are tested.

Lo, Watanabe, and Sugimura (2003) stated that ''analyses of GHF9 … resulted in trees with poorly resolved nodes (data not shown).'' In contrast, Libertini, Li, and McQueen-Mason (2004) were able to robustly resolve the relationships between plant GHF9 sequences, and Tokuda et al. (2004) achieved the same with termite sequences. We therefore addressed alignment and phylogenetic reconstruction with caution. One aim was to include as many sequences (both full length and partial) as possible, giving two main advantages: improving overall alignment and reducing problems associated with long-branch attraction (Felsenstein 1978). As mentioned, there are two families of the GHF9 gene (Tomme, Warren, and Gilkes 1995) and one has only been discovered in the bacteria (subgroup E1). As the two families are highly divergent in protein sequence, we were unable to include subgroup E1 in the analysis. The relationship between subgroups E1 and E2 therefore remains unresolved.

Full-length protein sequences were initially aligned using ClustalW (Thompson et al. 1997) and adjusted by eye. Partial sequences were added manually. The alignment of 316 GHF9 protein sequences is available in NEXUS format as Supporting Information. Prior to phylogenetic analysis, signal peptide sequences and other N-terminal extensions, gap-prone segments, and C-terminal extensions peculiar to individual taxa were excluded (N- and C-terminal extensions are common in GHF9 cellulases and

commonly comprise cellulose-binding domains or transmembrane anchor segments). In total, 436 characters were used for the phylogenetic analysis.

The amino acid sequences of this unambiguously aligned portion of the alignment were subjected to Bayesian, maximum likelihood, and neighbor-joining phylogeny reconstruction methods. Three different levels of analysis were carried out to enable a balance between adequate taxon sampling and speed of analysis. The first analysis included all full-length sequences and was used to identify and exclude nearly identical sequences. The second analysis was on the resulting reduced set of full-length sequences (most of the excluded sequences were plant GHF9 genes). In principle, maximum likelihood methods can allow for missing data, but there can still be problems (Kearney 2002; Philippe et al. 2004). As many of the sequences (especially from the Metazoa and primitive plants) were partial gene sequences from expressed sequence tags (ESTs), a final analysis was carried out including the reduced set of full-length sequences and all partial sequences.

With MrBayes v3.0b4, a mixed model of amino acid evolution was used with and without a gamma correction (4 categories of variable sites) (Huelsenbeck and Ronquist 2001). Four chains were run for a million generations. Prior to estimating support for the topology, we checked that the chains had converged and that the log likelihood was stationary. Neighbor-joining trees were constructed in PHYLIP v3.62 (Felsenstein 2004), using the JTT (Jones, Taylor, and Thornton 1992) amino acid substitution matrix. Finally, maximum likelihood analyses were carried out using Phyml v2.4 (Guindon and Gascuel 2003), again using the JTT amino acid substitution matrix. Support for the resulting neighbor-joining and maximum likelihood trees was assessed by bootstrap resampling, using routines within the same packages to produce extended majority rule consensus trees. As with MrBayes, for both neighbor-joining and maximum likelihood methods, we also allowed for rate variation between sites, and compared the resulting trees against the non–rate-corrected phylogenies.

The method of Shimodaira and Hasegawa (1999) was used to test the monophyly of the Metazoa, by comparing trees of different topology, and was implemented in PAML (Yang 1997). Specifically, we compared the difference in likelihood between the maximum likelihood tree (Metazoa = monophyletic) and that of a reduced topology tree (main branches in the Metazoa reduced to a polytomy with non-metazoan phyla).

Intron Positions

Although most metazoan GHF9 cellulases are only known from EST sequences, a few genomic sequences are available in public databases (e.g., AB019146, AB125892, AY176645). We compared the intron positions of metazoan GHF9 genes against representative taxa from the Viridiplantae, *Dictyostelium*, and Fungi. The GHF9 gene intron positions have been characterized for some metazoan taxa such as termites (Tokuda et al. 1999), and we were able to infer intron positions for other taxa (e.g., sea urchin) based on comparisons between ESTs and genomic sequence.

Results

New GHF9 Genes

We identified over 300 GHF9 genes in diverse eukaryotes, with a particular concentration in the Metazoa and Viridiplantae. For the first time, GHF9 cellulases were recognized in two new animal phyla, in Annelida (earthworm) and Echinodermata (sea urchin). In total, GHF9 cellulases were identified in 5 metazoan phyla, 10 classes, and 18 orders. The results are summarized in tables 2 and 3, with some important details below. Accession numbers of all sequences are in the supporting material.

From ESTs, we added previously unrecognized cellulases (see table 2) from arthropods, an annelid, mollusks, and an echinoderm. The pond snail *Lymnaea stagnalis* GHF9 gene was isolated during our own EST sequencing survey (Davison and Blaxter 2005). The cDNA clone corresponding to a *Biomphalaria glabrata* (Mollusca) GHF9-like EST was obtained from Anne Lockyer (Natural History Museum, London, United Kingdom) and completely sequenced (GenBank accession number AY651250). A GHF9 EST purportedly from *Schistosoma mansoni* (CD132744) is probably a contaminant because (1) the DNA sequence overlaps with a *B. glabrata* EST, (2) the *S. mansoni* tissue was extracted from a *B. glabrata* host, and (3) the partial ''*S. mansoni*'' sequence groups with *B. glabrata* sequences in phylogenies. A GHF9 gene from the sea urchin *Strongylocentrotus purpuratus* was isolated in an EST survey, though not characterized (Zhu et al. 2001). Two *Lumbricus rubellus* (Annelida) GHF9 genes were derived from our own study of earthworm gene expression (Blaxter, unpublished data).

Several putative GHF9 genes were also identified from genomic DNA sequences (tables 2 and 3), including the honeybee *A. mellifera*, sea squirts *C. intestinalis* and *C. savignyi*, and slime mold *D. discoideum*. *Dictyostelium discoideum* has at least 7 and possibly 11 GHF9 genes (Libertini, Li, and McQueen-Mason 2004). Three *C. savignyi* GHF9 genes were assembled from unannotated whole-genome shotgun sequence. An additional GHF9 gene from the sea urchin *S. purpuratus* was assembled from BAC-end sequences (see http://www.hgsc.bcm.tmc.edu). In addition to the new metazoan GHF9 genes, five fungal genomes, four basidiomycetes, and a chytridiomycete yielded one to two GHF9 genes each (table 3). However, none of the other complete fungal genomes (e.g., *Neurospora*, *Aspergillus*) were found to contain GHF9 genes. As expected, plant genomes yielded many GHF9 homologues: the fully sequenced genomes of *Arabidopsis thaliana* and *Oryza* sp. contain over 20 and 7 paralogues, respectively (Libertini, Li, and McQueen-Mason 2004) (see http://afmb.cnrs-mrs.fr/CAZY/index.html), and we identified additional unrecognized homologues in conifers (Kirst et al. 2003; Ujino-Ihara et al. 2003), cycads (Brenner et al. 2003; Brenner et al. unpublished GenBank submissions), a fern (Chatterjee et al. unpublished GenBank submissions), *Welwitschia* (gnetophyte; dePamphilis et al.

Table 2
Metazoan GHF9 Subgroup E2 Endo-Beta-1,4-Glucanases

Phylum	Class	Order	Species		No. genes (if >1)	Evidence
Annelida	Oligochaeta	Haplotaxida	*Lumbricus rubellus*	Earthworm	2	ESTs
Arthropoda	Malacostraca	Decapoda	*Cherax quadricarinatus*	Crayfish		cDNA; genomic DNA
			Homarus americanus	Lobster	2	ESTs
			Callinectes sapidus	Crab		EST
		Amphipoda	*Gammarus pulex*	Shrimp		ESTs
	Branchiopoda	Diplostraca	*Daphnia magna*	Water flea	2	ESTs
	Insecta	Hymenoptera	*Apis mellifera*	Honeybee		Genomic DNA
		Isoptera	*Coptotermes formosanus*	Termite		cDNA
			Mastotermes darwiniensis	Termite	2	cDNA
			Nasutitermes takasagoensis	Termite	2	Genomic DNA
			Reticulitermes speratus	Termite		Genomic DNA
		Blatteria	*Panesthia cribrata*	Cockroach		cDNA
		Coleoptera	*Timarcha balearica*	Beetle		EST
Chordata	Ascideacea	Enterogona	*Ciona intestinalis*	Sea squirt	9	Genomic DNA; ESTs
			Ciona savignyi	Sea squirt	3	Genomic DNA; ESTs
		Stolidobranchia	*Molgula tectiformis*	Sea squirt		EST
			Botryllus schlosseri	Sea squirt		EST
			Halocynthia roretzi	Sea squirt		EST
	Appendicularia	Appendiculariae	*Oikopleura dioica*	Sea squirt		Genomic DNA
	Mammalia	Primates	*Homo sapiens*	Human		cDNA
Echinodermata	Echinoidea	Echinoida	*Strongylocentrotus purpuratus*	Sea urchin	3	EST; BAC-end sequences
Mollusca	Gastropoda	Pulmonata	*Biomphalaria glabrata*	Bloodfluke planorb	2	ESTs
			Lymnaea stagnalis	Great pond snail		EST
		Vetigastropoda	*Haliotis discus*	Abalone		cDNA; genomic DNA
	Bivalvia	Veneroida	*Dreissena polymorpha*	Mussel		EST
		Pectinoida	*Argopecten irradians*	Bay scallop		EST
		Ostreoida	*Crassostrea virginica*	Oyster		ESTs

NOTE.—The human sequence may be a contaminant from an unknown metazoan.

unpublished GenBank submissions), and mosses (Nishiyama et al. 2003; Oliver et al. unpublished GenBank submissions) (table 3). In comparison, relatively few GHF9 genes (subgroup E2) were found in prokaryotes, even though over 150 complete genome sequences are available. Furthermore, while GHF9 (subgroup E2) cellulases are found in a relatively broad range of Eubacteria, the number of representatives per bacterial division is low (table 3).

Phylogenies

The phylogenies have a number of conspicuous features strongly supported by all methods. Each of the groups Eubacteria, Fungi, Amoebozoa, Viridiplantae, and Metazoa are monophyletic, with 100% support in Bayesian reconstructions (fig. 1). The same monophyletic groups are recovered using both maximum likelihood and neighbor-joining methods, with a single exception: the monophyly of the fungi is not supported in the maximum likelihood phylogeny because the Chytridiomycota (*Piromyces*) and Basidiomycota (*Cryptococcus*, *Ustilago*, and *Phanerochaete*) are separate. Bootstrapping of the neighbor-joining and maximum likelihood trees produced a consensus phylogeny with the same monophyletic groups, with this single exception, and methods accounting for between-site variation did not affect the topology. Using maximum likelihood and neighbor-joining methods, the monophyly of the Viridiplantae and Amoebozoa was very strongly supported, whereas there was generally somewhat lower support for the monophyly of the Eubacteria and Metazoa. The method of Shimodaira and Hasegawa (1999) provided additional evidence for the monophyly of the Metazoa: the difference in log likelihood between the best tree and a reduced topology tree was significant ($-\ln L = 41{,}337.49$, $41{,}365.22$; $P = 0.014$).

Therefore, as the null hypothesis was that GHF9 genes are related by vertical descent, the phylogeny (fig. 1) is entirely consistent with that, and provides no evidence to support the alternative hypothesis of horizontal gene transfer between kingdoms. The phylogenetic analysis does not resolve the relationship between different kingdoms, so the uncertainty about the relationship at the base of the tree is illustrated in figure 1 by a shaded region.

Within the monophyletic plant, animal, and bacterial groups, there is strong support for certain higher order groupings (e.g., Mollusca) but weaker support for the branches that describe the relationship between them (fig. 2). Again, this finding was confirmed using all three phylogenetic methods. In the Metazoa, genes that were isolated from species in the same phylum tend to group together (fig. 2*B*). In plants and bacteria, the presence of multiple paralogues from one species or genus is shown by their independent grouping within each kingdom (fig. 2*A* and *D*). Interestingly, and in keeping with accepted relationships within Viridiplantae, sequences from conifers and cycads tend to arise basally compared to their angiosperm orthologues, and Lilopsida (rice, lily, wheat) genes arise basally compared with orthologues from dicotyledon plants (fig. 2*A*).

Table 3
Alignment of Three Conserved Regions in GHF9 Subgroup E2 from Five Kingdoms, Including Taxa from Five Metazoan Phyla

Kingdom or Subkingdom	Phylum or Subgroup	Species		203 Conserved Region I * *	216 Conserved Region II *	661 Conserved Region III * * 689 6
Eubacteria	Actinobacteria	*Cellulomonas fimi*		LTGGWY**D**AG**D**HVKFGFP	PPTAP**H**HRTAHGS	NDAYT**D**SRQDYVAN**E**VAT
		Thermobifida fusca		LTGGWY**D**AG**D**HVKFGFP	PPRNP**H**HRTAHGS	NDAYT**D**DRQDYVAN**E**VAT
		Thermonospora sp.		LTGGWY**D**AG**D**HVKFGFP	PPRNP**H**HRTAHGS	NDAYT**D**DRQDYVAN**E**VAT
	Cyanobacteria	*Synechocystis* sp.		LTGGYH**D**AG**D**HGKFGLP	FPQQP**H**HRAASGV	NDSYN**D**SRDDYISN **E**VAI
	Firmicutes	*Anaerocellum thermophilum*		LTGGWF**D**AG**D**HVKFNLP	PPKRP**H**HRTAHSS	DDSYT**D**DISNYVNN **E**VAC
		Bacillus licheniformis		LTGGWY**D**AG**D**HVKFGLP	PPKHP**H**HRTAHGS	DDSYR**D**DITDYASN **E**VAI
		Caldicellulosiruptor saccharolyticus		LTGGWF**D**AG**D**HVKFNLP	PPKRP**H**HRTAHSS	DDSYT**D**DISNYVNN **E**VAC
		Caldicellulosiruptor sp.		LTGGWH**D**AG**D**HVKFNLP	YPQHP**H**HRNAHSS	DDSYN**D**DITDYVQN **E**VAC
		Clostridium acetobutylicum		LTGGWY**D**AG**D**HVKFNLP	PPEHP**H**HRTAESS	NDKFD**D**DVANFNQN**E**PAC
	Proteobacteria	*Myxobacter* sp.		LTGGWY**D**AG**D**HVKFGLP	PPKHP**H**HRTAHGS	DDSYR**D**ETNDYVSN**E**VAI
Fungi	Chytridiomycota	*Piromyces* sp.		LTGGYY**D**AG**D**NVKFNFP	SPKAV**H**HRAASGT	KDEYT**D**SRKNYEMN**E**VAL
	Basidiomycota	*Phanerochaete chrysosporium*		LSGGYY**D**AG**D**YI KYTFP	APSNP**H**SALATGA	DDLFW**D**LRSDWVES**E**VGL
		Ustilago maydis		LSGGYY**D**AG**D**YI KATYP	SPQNP**H**SA MASGG	QDRFF **D**IRDDWPQT**E**IAL
		Cryptococcus neoformans		LVGGWY**D**AG**D**YI KATFP	SPSNP**H**SAPASGG	SDQFW**D**WRDDWVQT**E**IAL
Amoebozoa		*Dictyostelium discoideum*	Slime mold	LSGGYF **D**AG**D**GVKFGLP	YPINP**H**HRAAHHS	NDEYT**D**DRTDYISN **E**VAT
Metazoa	Annelida	*Lumbricus rubellus*	Earthworm	LTGGWY**D**AG**D**HVKFGFP	PPQRP**H**HRSSSCP	NDNYE**D**VRSDYISN **E**VAT
	Arthropoda	*Apis mellifera*	Honeybee	LTGGYY**D**AG**D**FV KFGFT	PPKQP**H**HAASSCP	ADKFH**D**HREDYVYT**E**VTL
		Cherax quadricarinatus	Crayfish	LTGGYY**D**AG**D**HVKFGFP	PPTRP**H**HRS SSCP	DGSYN **D**DRQDYQHN**E**VAC
		Daphnia magna	Water flea	???????????????	PPVKC**H**HRGASCP	NDYYN**D**ARDDYRSN**E**VAL
		Homarus americanus	Lobster	LTGGYY**D**AG**D**HVKFGFP	???????????????	DDSYV **D**DRSDYVHN**E**GAC
		Coptotermes formosanus	Termite	LTGGYY**D**AG**D**FV KFGFP	PPVRP**H**HRSSSCP	NDSYT **D**SRSDYISN **E**VAT
		Gammarus pulex	Shrimp	???????????????	PPTKP**H**HRSSSCP	NDGYV **D**DRNDYVHN**E**VAC
		Mastotermes darwiniensis	Termite	LTGGYY**D**AG**D**YVKFGFP	PPTHP**H**HRSSSCP	NDNYE**D**LRSDYVAN**E**VAT
		Nasutitermes takasagoensis	Termite	LTGGYF **D**AG**D**FV KFGFP	PPTRP**H**HRSSSCP	NDNYV **D**DRSDYVHN**E**VAT
		Nasutitermes walkeri	Termite	LTGGYF **D**AG**D**FV KFGFP	PPTRP**H**HRSSSCP	NDNYV **D**DRSDYVHN**E**VAT
		Reticulitermes speratus	Termite	LTGGYY**D**AG**D**FV KFGFP	PPVRP**H**HRSSSCP	NDSYT **D**ARSDYIS N**E**VAT
		Panesthia cribrata	Cockroach	LTGGYY**D**AG**D**FV KFGFP	YPTHE**S**HRSSSCP	NDDYE **D**LRSDYVHN**E**VAD
	Chordata	*Botryllus schlosseri*	Sea squirt	???????????????	???????????????	DGSYV**Y**DRGDYVKN**E**IAT
		Ciona intestinalis	Sea squirt	LTGGWY**D**AG**D**NI KFGFP	SPQKP**H**HRASSCP	NGAYT**D**DRSDYISN **E**VAT
		Ciona intestinalis	Sea squirt	LSGGWY**N**GG**G**AVKTTSL	APQNP**H**HRSSSCG	FGRYV **D**RASDYIRN**E**VAI
		Ciona intestinalis	Sea squirt	LSGGYY **V**SG**D**YV KYGFP	YPTQP**H**HRASFLI	TDQFT **N**DRSDYRSN**G**VIS
		Ciona intestinalis	Sea squirt	LSGGYF **T**DG**G**FV KYGFP	SPDRP**Y**HRASSCP	WDAFS**N**VRSDTKHN**S**VSI
		Ciona savignyi	Sea squirt	LSGGYY **D**AG**D**NVKFGFP	SPQRP**H**HRA-SCP	SGAYT **D**DRSDYISN**E**VAT
		Halocynthia roretzi	Sea squirt	???????????????	???????????????	TGIYL **D**NRAIVEQ S**E**VAC
		Homo sapiens	Human	LSRGWY **E**AA**N**TMKWGLP	YPSKP**Y**HKSSYNS	DDTWY**D**DRSNYEYS**E**VTQ
		Molgula tectiformis	Sea squirt	???????????????	YPRRS**Y**HKASSCP	WDNYE**D**I R-DPKFN**K**VGI
		Oikopleura dioica	Sea squirt	LSGGYY **D**GGG**G**FIKYNFP	FPTRY**Y**HKESFCP	DRYDDD**D**PWKQQQSS**V**AMD
	Echinodermata	*Strongylocentrotus purpuratus*	Sea urchin	LTGGWY **D**AG**D**HVKFGLP	PPLRH**P**TYCSSCP	YDHYN**D**DRGDYISN**E**VAC
	Mollusca	*Biomphalaria glabrata*	Planorb	LTGGWY**D**AG**D**HVKFNFP	YPLRP**H**HRASS??	TMTYT**D**DRSNYVNN**E**VAC
		Crassostrea virginica	Oyster	???????????????	YPRQP**H**HRGSSCP	???????????????
		Lymnaea stagnalis	Pond snail	VVGGWH**D**AG**D**HVKFQLP	YPKNP**H**HRASSCP	NDNYE**D**KRSDYIK N**E**VAL
		Haliotis discus	Abalone	LTGGWY**D**AG**D**HVKFSLP	YPRNP**H**HRSASCP	DDSYK**D**NREDYVHN**E**VAC
		Argopecten irradians	Bay scallop	LTGGWY**D**AG**D**LVKFNFP	???????????????	???????????????
		Dreissena polymorpha	Mussel	???????????????	???????????????	DDSYK**D**DRNDYVKN**E**VAA

Table 3
Continued

Kingdom or Subkingdom	Phylum or Subgroup	Species		203 Conserved Region I 219 * *	261 Conserved Region II 268 *	269 Conserved Region III 368 * *
Viridiplantae	Tracheophytes (angiosperms)	*Hordeum vulgare*	Barley	L V G G F Y **D** A G **D** A I K F N F P	Y P K R V **H** H R G A S I P	H D G F K **D** I R T N Y N Y T **E** P T L
		Lilium longiflorum	Lily	L T G G Y Y **D** A G **D** N VK F G F P	F P L H I **H** H R G S S I P	N D S F A **D** DR D N Y S Q S **E** P A T
		Oryza sativa	Rice	L V G G F Y **D** A G **D** A I K F N Y P	Y P K R A **H** H R G A S I P	H D G F K **D** VR T N Y N Y T **E** P T L
		Triticum aestivum	Wheat	L V G G F Y **D** A G **D** A I K F N Y P	Y P K R V **H** H R G A S I P	H D G F K **D** I R T N Y N Y T **E** P T L
		Arabidopsis thaliana	Thale cress	L S K G L Y **D** A G **D** H MK F G F P	Y P E F V **H** H R G A S I P	N D T F I **D** A R N N S M Q N **E** P S T
		Atriplex lentiformis	Saltbush	L V G G Y Y **D** A G **D** N VK F G L P	Y P Q H I **H** H R A S S LP	G D K F T **D** DR N N Y R Q S **E** P A T
		Brassica napus	Rape	L V G G Y Y **D** A G **D** A I K F N F P	Y P K H V **H** H R G A S I P	R D G F R **D** VR T N Y N Y T **E** P T L
		Capsicum annuum	Pepper	L V G G Y Y **D** A G **D** N VK F G L P	Y P L R V **H** H R G S S LP	R D N F E **D** DR N N Y Q Q S **E** P A T
		Citrus sinensis	Orange	L T G G Y Y **D** A G **D** N VK F N F P	F P R R I **H** H R G S S LP	N D G F P **D** DR S D Y S H S **E** P A T
		Fragaria × ananassa	Strawberry	L T G G Y Y **D** A G **D** N VK F G F P	Y P Q R I **H** H R G S S LP	S D A F P **D** S R P Y F Q E S **E** P T T
		Lycopersicon esculentum	Tomato	L V G G Y Y **D** A G **D** N VK F G L P	Y P R Q V **H** H R A S S I V	Y D N F A **D** Q R D N Y E Q T **E** P A T
		Malus × domestica	Apple	L T G G Y Y **D** A G **D** N VK F N FP	Y P K R I **H** H R G S S LP	N D G F P **D** D R G D Y S H S **E** P A T
		Medicago truncatula	Clover	L I G G Y Y **D** S G **N** N I K F T FT	F P V Q V **H** H R S A S I P	N D H F T **D** Q R S N K R F T **E** P T I
		Nicotiana alata	Tobacco	L T G G Y Y **D** A G **D** N VK F G FP	Y P Q K I **H** H R G A S I V	S D N Y N **D** S R T N F Q Q A **E** A AT
		Persea americana	Avocado	L V G G Y Y **D** A G **D** N LK F G L P	Y P Q H V **H** H R G S S LP	R D S F S **D** DR N N Y Q Q S **E** P A T
		Phaseolus vulgaris	Bean	L I G G Y Y **D** A G **D** N VK F G WP	Y P K Q L **H** H R G S S I P	N D R F N **D** A R S D Y S H A **E** P T T
		Pisum sativum	Pea	L V G G Y Y **D** A G **D** N VK F G F P	Y P Q R I **H** H R G S S LP	H D R F P **D** Q R S D Y E Q S **E** P A T
		Populus alba	Poplar	L V G G Y Y **D** A G **D** N VK F G L P	Y P Q H V **H** H R G S S VP	R D N F A **D** D R N N Y Q Q S **E** P A T
		Prunus persica	Peach	L V G G Y Y **D** A G **D** N VK F G LP	Y P L H I **H** H R G S S LP	K D S F S **D** D R N N Y Q Q S **E** P A T
		Pyrus communis	Pear	L A G G F Y **D** A G **D** A I K F N F P	Y P K H V **H** H R G A S I P	H D G F R **D** V R S N Y N Y T **E** P T L
		Sambucus nigra	Elder	L T G G Y Y **D** A G **D** N VK F GW P	Y P L Q L **H** H R G A S I P	N D Q F N **D** V R S D Y S H L **E** T TT
	(ferns)	*Ceratopteris richardii*	Fern	L S G G M Y **D** ? ? ? ? ? ? ? ? ? ?	? ? ? ? ? ? ? ? ? ? ? ? ?	? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ?
	(coniferales)	*Pinus radiata*	Pine	L T G G Y Y **D** A G **D** N VK F G F P	F P E R I **H** H R G S S L P	N D H F S **D** E R N D Y A H S **E** P T T
		Cryptomeria japonica	Cedar	L T G G Y Y **D** A G **D** N VK F G F P	? ? ? ? ? ? ? ? ? ? ? ? ?	? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ?
	(cycads)	*Cycas rumphii*	Cycad	L V G G Y Y **D** A G **D** N MK F G F P	Y P R Q V **H** H R A S S I V	N D N F A **D** Q R D N Y E Q T **E** P A T
		Zamia furfuracea	Cycad	L V G G Y Y **D** A S **D** N MK F G F P	? ? ? ? ? ? ? ? ? ? ? ? ?	? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ?
	(gnetophytes)	*Welwitschia mirabilis*	Tree tumbo	L S K G L Y **D** A G **D** H I K F G L P	Y P R Q V **H** H R A S S I V	N D N F A **D** E R D N Y E Q T **E** P A T
	(Iycophytes)	*Selaginella lepidophylla*	Club moss	? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ?	Y P K H V **H** H R A A S I P	K D R F H **D** V R T N Y N Y T **E** P T V
	Bryophytes	*Physcomitrella patens*	Moss	L V G G Y Y **D** A G **D** N VK F G L P	Y P Q K L **H** H R G A S I P	N E T Y S **D** T R D N I L Q N **E** A S T
	(mosses)	*Tortula ruralis*	Moss	? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ?	Y P K F L **H** H R G A S I P	N E T Y T **D** S R V N I Q Q N **E** A S V

NOTE.—Catalytically important residues (indicated by an asterisk) are nearly always conserved, except in some deuterostomes (- = gap; ? = missing data). The human sequence may be a contaminant from an unknown metazoan.

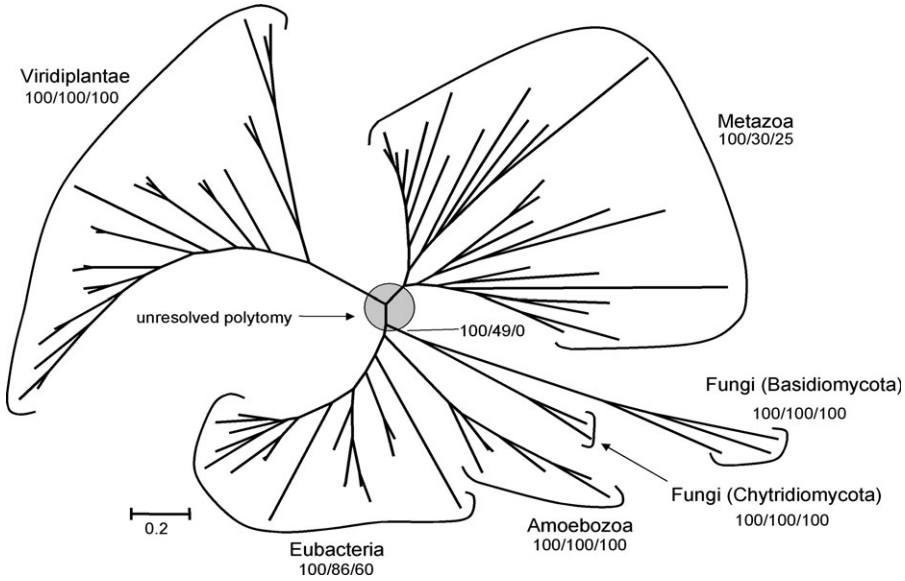

FIG. 1.—The diversity of GHF9 cellulases. This unrooted phylogram shows the topology supported by Bayesian analysis with a gamma correction. Metazoa, Viridiplantae, Fungi, Amoebozoa, and Eubacteria are all monophyletic, with 100% support. Maximum likelihood and neighbor-joining analyses concur with this, albeit with lower bootstrap support, except that the maximum likelihood does not support a monophyletic fungal group. The base of the tree (shaded) is unresolved. Support using Bayesian–neighbor-joining–maximum likelihood methods is shown.

As expected, phylogenetic analyses which included the partial EST sequences were compromised by missing data (some were only 25% of the full-length sequence alignment). Bootstrap support was low, and parts of the topology differ between methods (not shown). Nonetheless, the kingdom-level groups are still monophyletic in a Bayesian analysis (not shown). As in the full-length analysis (fig. 2), most of the new metazoan sequences are grouped with a gene sampled from the same phylum, whereas the more ''primitive'' plant sequences, from mosses, a fern, conifers, cycads, and *Welwitschia* (gnetophyte), tend to fall basally (supplementary fig. 1).

''Human'' GHF9

We identified a putative human GHF9 gene in EST data from a full-length heart cDNA library (Imanishi et al. 2004; FLJ38599). However, a corresponding sequence is not present in the draft human genome sequence (build 35), nor could we find a homologue in other vertebrates. One possibility is that this EST is a laboratory or informatic contaminant. However, we did not find significant numbers of nonhuman sequences in the other ~30,000 ESTs derived and submitted from the same project (Imanishi et al. 2004). Obvious contaminants were either vector sequences or clearly derived from Drosophila (e.g., AK094453, AK130952–AK130956). It remains possible that the gene has not been found in humans because it is rarely expressed or is located in a region that is difficult to clone (e.g., heterochromatin). However, we were unable to polymerase chain reaction amplify specific gene fragments from human genomic DNA, and in situ hybridization experiments with human chromosomes also failed (W. Bickmore, personal communication).

Introns

Three intron positions are conserved between taxa from three metazoan phyla and at least two are also shared with an echinoderm GHF9 genomic sequence (table 4). An *Arabidopsis* and a rice GHF9 gene (CAB45061, or NP_194157, AC137547) share one intron position with metazoan GHF9 genes (table 4).

Functional Site Analysis

A consensus of functionally important sites has been identified in previous studies of GHF9 gene action (Khademi et al. 2002; Lo, Watanabe, and Sugimura 2003; Suzuki, Ojima, and Nishita 2003). We compared the sequence of the newly identified GHF9 genes at these sites and to a core region surrounding the active site (table 3). Most sequences have the expected conserved amino acid at each of the five sites, except for some of the deuterostome sequences (table 3).

Discussion

Previously, Lo, Watanabe, and Sugimura (2003) used intron positional evidence to argue that GHF9 subgroup E2 genes from termites, abalone, and sea squirt have an ancient, common origin. Our results are entirely consistent with theirs and significantly extend this model: GHF9 genes are present in at least five metazoan phyla, and the monophyly of the metazoan GHF9 genes in the phylogeny suggests a single, ancient origin (fig. 1). Evidence from two further conserved intron positions also supports this conclusion (table 4).

GHF9 genes from the Viridiplantae and Metazoa are monophyletic, with high support, which is suggestive of an origin for the gene in an ancient eukaryote (fig. 1).

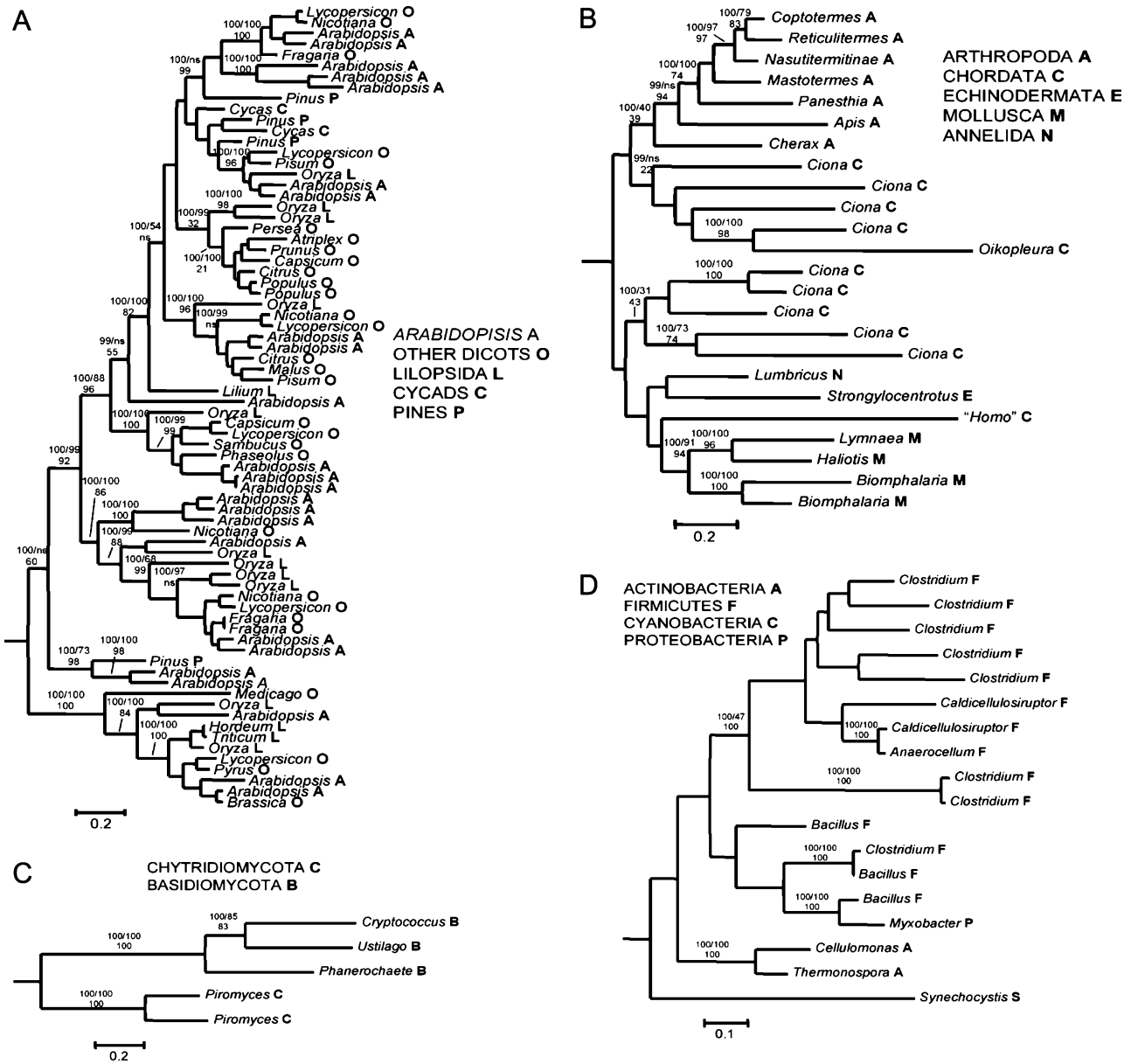

FIG. 2.—Kingdom-level analyses of GHF9 phylogeny. These rooted subtrees show in detail the within-kingdom relationships of GHF9 genes from (*A*) Viridiplantae, (*B*) Metazoa, (*C*) Fungi, and (*D*) Eubacteria using MrBayes with a gamma correction. The same general pattern was recovered using both neighbor-joining and maximum likelihood methods. In the plants (*A*), primitive members tend to arise basally to angiosperm representatives. In animals (*B*), GHF9 genes that were isolated from the same phylum tend to group together. In the bacteria (*D*), some sequences from distantly related bacteria tend to fall together. Branches with 99% or more support in the Bayesian tree are shown, followed by the support using neighbor-joining–maximum likelihood methods (ns = not supported).

However, GHF9 has been duplicated in some lineages (e.g., *Ciona*, *Biomphalaria*, *Arabidopsis*, *Oryza*) and lost in others (e.g., the completely sequenced *Anopheles gambiae*, *Drosophila melanogaster*, *Caenorhabditis elegans*, *Saccharomyces pombe*, and *Saccharomyces cerevisiae* genomes, as well as most complete fungal genomes). Within the Viridiplantae in particular (fig. 2*A*, supplementary fig. 1*A*), but also the Metazoa (fig. 2*B*, supplementary fig. 1*B*), multiple paralogues were identified from some taxa. In consequence, it is not surprising that the gene trees do not match accepted organismal phylogenies. However, the multiple paralogues found in fully sequenced plant genomes appear to be ancient as, in general, each paralogue group contains both primitive and angiosperm members, with mosses and ferns, conifers, cycads, and *Welwitschia* (gnetophyte) representatives tending to arise basally to angiosperm representatives (fig. 2*A*, supplementary fig. 1*A*). In the Metazoa, the arthropod sequences and the mollusk sequences were monophyletic (fig. 2*B*), with the sea squirt sequences clustering in two or more separate groups. The putative human GHF9 gene was robustly placed in the metazoan clade but not associated with other deuterostomes (fig. 2*B*). It seems likely that this sequence is a contaminant from an unknown metazoan (Imanishi et al. 2004).

Table 4
Conserved Introns in Four Metazoan Phyla, One of Which Is Also Shared with *Arabidopsis thaliana* and *Oryza sativa*

Phylum	Species		Intron 1 (frame = 1)	Intron 2 (frame = 0)	Intron 3 (frame = 2)
Arthropoda	*Nasutitermes takasagoensis*	Termite	Y F D A G D TAC TTT GAC G/CT GGT GAT	G V Q V L L GGT GTA CAG/GTT CTA CTG	H R S S S C CAT AGA TCT AG/T TCG TGC
Arthropoda	*Cherax quadricarinatus*	Crayfish	W Y - - - - TAC TAC GAT G/CT GGT GAT	A C - L - - GGA GTA CAG/GCA CTC CTC	- - - A - - CAC CGT TCC AG/C TCG TGC
Mollusca	*Haliotis discus*	Abalone	W Y - - - - TGG TAT GAT G/CT GGT GAC	- - - Q M GCT TGC CAG/CTG CTC CTG	CAC AGA TCC GC/C TCC TGC
Echinodermata	*Strongylocentrotus purpuratus*	Sea urchin	Y Y - - - - TGG TAT GAT G/CT GGA GAC	- - - A - - GGT GTA CAG/CAG ATG TTA	
Chordata	*Ciona intestinalis*	Sea squirt	W Y - - - - TGG TAC GAC G/CT GGA GAC	- - - - M GGA GTT CAG/GTC CTC ATG	- - - A - - - CAC AGA GCC AG/T TCT TGC
Tracheophytes	*Arabidopsis thaliana*	Thale cress		- A - L - - GGG GCT CAG/TTG CTG TTA	
Tracheophytes	*Oryza sativa*	Rice		- A - - - - GGT GCA CAG/GTT CTT TTG	

NOTE.—The intron position is shown by "/". The amino acid sequences surrounding introns 1 and 3 correspond to the conserved regions I and II in table 3. Amino acids are only shown when they differ from the corresponding amino acid in *Nasutitermes takasagoensis*. There are presently insufficient assembled contigs to determine whether intron 3 is also conserved in the sea urchin. The genomic sequences used are ci0100130070 (sea squirt; see http://genome.jgi-psf.org/ciona4), AB019146 (termite), AB125892 (abalone), AY176645 (crayfish), contigs 310112/contig617229 (sea urchin; see http://www.hgsc.bcm.tmc.edu/projects/seaurchin/), CAB45061/NP_194157 (*A. thaliana*), and AC137547 (*O. sativa*).

In theory, an independent approach to investigate the evidence for horizontal gene transfer would be to identify genes that appear atypical in their current genomic context, for which one explanation is that they were introduced recently from a foreign source. However, the methods are only suitable when the gene was acquired recently, because over time the foreign genes will come to resemble the host genes (Ochman, Lawrence, and Groisman 2000). There is also evidence to suggest that successful horizontal transfer requires codon usage compatibility between foreign genes and recipient genomes (Medrano-Soto et al. 2004), which would make this method even less sensitive. The same approach could also be used to identify contaminating ESTs, such as the possible human GHF9 sequence, but there are significant problems with applying the method to nonbacterial systems. For instance, codon usage bias is not an effective measure where selection for translational efficiency is weak (e.g., in humans), so the greater effect on codon usage is provided by differences in GC content, which are in turn determined by genomic position (Plotkin, Robins, and Levine 2004).

It is interesting that there is still a widespread belief, especially in the popular science literature (for example see Chap. 11, p. 311, and Notes, p. 443, in Morris 2003), that animals do not have endogenous cellulase genes. In fact, a number of early works claim to demonstrate endogenous cellulase activity (see references within Watanabe and Tokuda [2001] and Yokoe and Yasumasu [1964]). While discrimination between glycosyl hydrolase families was not possible at the time, in a most comprehensive study Yokoe and Yasumasu (1964) found evidence for cellulases in annelids, arthropods, mollusks, echinoderms, and chordates, which exactly mirrors our own findings. Intriguingly, the same study also found evidence for cellulases in *Urechis* (phylum: Echiura), *Lingula* (Brachiopoda), and *Physcosoma* (Sipunculida). Presently, few DNA sequences are available from any of these phyla, so it will be interesting to look for cellulase genes when sufficient ESTs have been isolated.

Functional and crystal structure analyses of prokaryotic and eukaryotic GHF9 cellulases, including a termite cellulase (Khademi et al. 2002), have defined key structural and catalytic residues. The majority of the putative GHF9 homologues we identified appear to be true cellulases because five key catalytic residues are almost invariant across the Eukaryota (table 3). Strikingly, in deuterostome GHF9 genes, especially those from sea squirts, there are frequent substitutions in these critical residues, some nonconservative (table 3). Thus, these cellulases may have altered cellulolytic activity or even gained new functions. This idea has a precedent in the discovery of vertebrate ''chitinases,'' which actually have a role in an innate immune recognition system (Zhu et al. 2004). It is already clear that the role of GHF9 may switch during metazoan evolution: termites express endogenous cellulases in different organs, depending upon the species and whether hypermastigote symbionts are present or not (Tokuda et al. 2004).

The monophyly and diversity of eukaryotic GHF9 cellulases, and corresponding paucity in prokaryotes, in conjunction with the conserved intron position between plants and animals, is suggestive evidence for the presence of an

ancestral GHF9 cellulase gene in an early eukaryote, predating the divergence between eukaryotic kingdoms. If plants, animals, *Dictyostelium*, and fungi had independently gained GHF9 by horizontal gene transfer, then prokaryote GHF9 genes would disrupt the monophyly of each group. In particular, the monophyly of the metazoan GHF9 genes provides compelling evidence for their ancient and common origin in animals, predating the divergence of the five phyla. In contrast, in Eubacteria the close relationship between some sequences from different phyla (e.g., *Bacillus* and *Myxobacter*) is most easily explained by horizontal gene transfer within Eubacteria (Ochman, Lawrence, and Groisman 2000).

Horizontal gene transfer is a significant feature of genome evolution in prokaryotes, but the relevance to eukaryotic evolution is much more uncertain (Ochman, Lawrence, and Groisman 2000; Genereux and Logsdon 2003). Most previously reported cases of prokaryote to eukaryote horizontal gene transfer have subsequently been falsified (Salzberg et al. 2001; Stanhope et al. 2001), and this now seems to be the case for GHF9. A few exceptional incidences of horizontal gene transfer have been identified, in addition to one previously mentioned (Smant et al. 1998), including the transfer of a *Wolbachia* genome segment to the insect host (Kondo et al. 2002), of multiple genes to diplomonads (Andersson et al. 2003), and between a protist and a cnidarian (Steele et al. 2004). As present-day eukaryote GHF9 genes are probably derived from an ancient eukaryote gene, instead of by horizontal gene transfer, many lineages must have lost the gene. Similar patterns of lineage-specific loss have been described for other genes, such as soluble adenylyl cyclase, which is present in vertebrates but has been lost in Drosophila, *Caenorhabditis*, *Arabidopsis*, and *Saccharomyces* (Roelofs and Van Haastert 2002).

In summary, we report evidence for an ancient and widespread eukaryotic endoglucanase with many metazoan representatives. It is intriguing that the last common ancestor of all deuterostomes was probably able to directly digest, or even synthesize cellulose using endogenous genes, as do sea squirts today. The lack of GHF9 in the genomes of many animal models (e.g., fly, mosquito, nematode) underlines the prevalence of gene loss in evolution and shows that reliable inference of gene evolution requires adequate taxon sampling. As most bilaterian phyla have so far been neglected in sequencing surveys (Blaxter 2002), we predict that many more eukaryotic cellulases, especially GHF9 genes, will be discovered as genome projects broaden their taxonomic spread. We expect additional protist taxa to have cellulases, as ciliates and flagellates are common gut commensals implicated in cellulose digestion, but whether these activities derive from GHF9-type enzymes is not currently known. At least some protists, such as the hypermastigote commensals of termites (Ohtoko et al. 2000; Li et al. 2003), have GHF7 and GHF45 genes. Finally, the additional biochemical functionality of Metazoa inferred from genome surveys suggests that as the phylogenetic scope of sequencing increases, additional biosynthetic and degradative capabilities may be found which are lacking in model organisms.

Supplementary Material

Accession numbers of sequences used in this study. Supplementary figure 1. Color versions of figures 1 and 2 for online version of manuscript.

SUPPLEMENTARY FIG. 1.—Kingdom-level analyses of GHF9 phylogeny, including all partial EST sequences. These rooted subtrees illustrate the pattern of within-kingdom relationships of GHF9 genes from (*A*) Viridiplantae and (*B*) Metazoa, using MrBayes with a gamma correction. They do not show the definitive relationships because the use of partial sequences, with missing data, resulted in lower support for some of these branches, and some rearrangements compared with the full-length analysis (e.g., *Apis mellifera*, bee, moves out of Arthropoda to cluster with the partial *Timarcha balearica*, beetle, sequence). In the plants (*A*), primitive members tend to arise basally to angiosperm representatives. In the Metazoa (*B*), GHF9 genes that were isolated from the same phylum tend to group together. * = 100% support; # = 95%–99% support using MrBayes.

Acknowledgments

We thank Wendy Bickmore, Shelagh Boyle, Anne Lockyer, Ann Hedley, Liz Bailes, and Ralf Schmid for support and analysis and Katelyn Fenn and Ralf Schmid for comments on the manuscript. Two anonymous referees gave useful comments. Funding was provided by the Royal Society (A.D.) and NERC (M.B.).

Literature Cited

Andersson, J. O., A. M. Sjogren, L. A. M. Davis, T. M. Embley, and A. J. Roger. 2003. Phylogenetic analyses of diplomonad genes reveal frequent lateral gene transfers affecting eukaryotes. Curr. Biol. **13**:94–104.

Blaxter, M. L. 2002. Genome sequencing: time to widen our horizons. Brief. Funct. Genomics and Proteomic. **1**:7–9.

Brenner, E. D., D. W. Stevenson, R. W. McCombie et al. (13 co-authors). 2003. Expressed sequence tag analysis in *Cycas*, the most primitive living seed plant. Genome Biol. **4**:R78.

Davison, A., and M. L. Blaxter. 2005. An expressed sequence tag survey of gene expression in the pond snail *Lymnaea stagnalis*, an intermediate vector of *Fasciola hepatica*. Parasitology (in press).

Dehal, P., Y. Satou, R. K. Campbell et al. (84 co-authors). 2002. The draft genome of Ciona intestinalis: insights into chordate and vertebrate origins. Science **298**:2157–2167.

Eigenheer, A. L., C. I. Keeling, S. Young, and C. Tittiger. 2003. Comparison of gene representation in midguts from two phytophagus insects, *Bombyx mori* and *Ips pini*, using expressed sequence tags. Gene **316**:127–136.

Felsenstein, J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. Syst. Zool. **27**:401–410.

———. 2004. PHYLIP (phylogeny inference package). Version 3.6. Distributed by the author, Department of Genome Sciences, University of Washington, Seattle.

Genereux, D. P., and J. M. Logsdon. 2003. Much ado about bacteria-to-vertebrate lateral gene transfer. Trends Genet. **19**:191–195.

Girard, C., and L. Jouanin. 1999. Molecular cloning of a gut-specific chitinase cDNA from the beetle *Phaedon cochleariae*. Insect Biochem. Mol. Biol. **29**:549–556.

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

Harada, Y., Y. Hosoiri, and R. Kuroda. 2004. Isolation and evaluation of dextral-specific and dextral-enriched cDNA clones as candidates for the handedness-determining gene in a freshwater gastropod, *Lymnaea stagnalis*. Dev. Genes Evol. **214**:159–169.

Henrissat, B. 1991. A classification of glycosyl hydrolases based on amino acid sequence similarities. Biochem. J. **280**:309–316.

Huelsenbeck, J. P., and F. Ronquist. 2001. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics **17**:754–755.

Imanishi, T., T. Itoh, Y. Suzuki et al. (154 co-authors). 2004. Integrative annotation of 21,037 human genes validated by full-length cDNA clones. PloS Biol. **2**:856–875.

Jones, D. T., W. R. Taylor, and J. M. Thornton. 1992. The rapid generation of mutation data matrices from protein sequences. Comput. Appl. Biosci. **8**:275–282.

Kearney, M. 2002. Fragmentary taxa, missing data, and ambiguity: mistaken assumptions and conclusions. Syst. Biol. **51**:369–381.

Khademi, S., L. A. Guarino, H. Watanabe, G. Tokuda, and E. F. Meyer. 2002. Structure of an endoglucanase from termite, *Nasutitermes takasagoensis*. Acta Crystallogr. D **58**:653–659.

Kikuchi, T. J., T. Jones, T. Aikawa, H. Kosaka, and N. Ogura. 2004. A family of glycosyl hydrolase family 45 cellulases from the pine wood nematode *Bursaphelenchus xylophilus* FEBS Lett. **572**:201–205.

Kirst, M., A. F. Johnson, C. Baucom, E. Ulrich, K. Hubbard, R. Staggs, C. Paule, E. Retzel, R. Whetten, and R. Sederoff. 2003. Apparent homology of expressed genes from wood-forming tissues of loblolly pine (*Pinus taeda* L.) with *Arabidopsis thaliana*. Proc. Natl. Acad. Sci. USA **100**:7383–7388.

Kondo, N., N. Nikoh, N. Ijichi, M. Shimada, and T. Fukatsu. 2002. Genome fragment of *Wolbachia* endosymbiont transferred to X chromosome of host insect. Proc. Natl. Acad. Sci. USA **99**:14280–14285.

Lee S. J., S. R. Kim, H. J. Yoon, I. Kim, K. S. Lee, Y. H. Je, S. M. Lee, S. J. Seo, H. D. Sohn, and B. R. Jin. 2004. cDNA cloning, expression, and enzymatic activity of a cellulase from the mulberry longicorn beetle, *Apriona germari*. Comp. Biochem. Physiol. B Biochem. Mol. Biol. **139**:107–116.

Li, L., J. Frohlich, P. Pfeiffer, and H. Konig. 2003. Termite gut symbiotic archaezoa are becoming living metabolic fossils. Eukaryot. Cell **2**:1091–1098.

Libertini, E., Y. Li, and S. J. McQueen-Mason. 2004. Phylogenetic analysis of the plant endo-beta-1,4-glucanase gene family. J. Mol. Evol. **58**:506–515.

Lo, N., H. Watanabe, and M. Sugimura. 2003. Evidence for the presence of a cellulase gene in the last common ancestor of bilaterian animals. Proc. R. Soc. Lond. B Biol. Sci. **270**:S69–S72.

Matthysse, A. G., K. Deschet, M. Williams, M. Marry, A. R. White, and W. C. Smith. 2004. A functional cellulose synthase from ascidian epidermis. Proc. Natl. Acad. Sci. USA **101**:986–991.

Medrano-Soto, A., G. Moreno-Hagelsieb, P. Vinuesa, J. A. Christen, and J. Collado-Vides. 2004. Successful lateral transfer requires codon usage compatibility between foreign genes and recipient genomes. Mol. Biol. Evol. **21**:1884–1894.

Morris, S. C. 2003. Life's solution: inevitable humans in a lonely universe. Cambridge University Press, Cambridge, United Kingdom.

Nakashima, K., L. Yamada, Y. Satou, J. Azuma, and N. Satoh. 2004. The evolutionary origin of animal cellulose synthase. Dev. Genes Evol. **214**:81–88.

Nishiyama, T., T. Fujita, T. Shin-I et al. (12 co-authors). 2003. Comparative genomics of *Physcomitrella patens* gametophytic transcriptome and *Arabidopsis thaliana*: implication for land plant evolution. Proc. Natl. Acad. Sci. USA **100**:8007–8012.

Ochman, H., J. G. Lawrence, and E. A. Groisman. 2000. Lateral gene transfer and the nature of bacterial innovation. Nature **405**:299–304.

Ohtoko, K., M. Ohkuma, S. Moriya, T. Inoue, R. Usami, and T. Kudo. 2000. Diverse genes of cellulase homologues of glycosyl hydrolase family 45 from the symbiotic protists in the hindgut of the termite *Reticulitermes speratus*. Extremophiles **4**:343–349.

Pennisi, E. 2002. Tunicate genome shows a little backbone. Science **298**:2111–2112.

Philippe, H., E. A. Snell, E. Bapteste, P. Lopez, P. W. H. Holland, and D. Casane. 2004. Phylogenomics of eukaryotes: impact of missing data on large alignments. Mol. Biol. Evol. **21**:1740–1752.

Plotkin, J. B., H. Robins, and A. J. Levine. 2004. Tissue-specific codon usage and the expression of human genes. Proc. Natl. Acad. Sci. USA **101**:12588–12591.

Roelofs, J., and P. J. M. Van Haastert. 2002. Deducing the origin of soluble adenylyl cyclase, a gene lost in multiple lineages. Mol. Biol. Evol. **19**:2239–2246.

Salzberg, S. L., O. White, J. Peterson, and J. A. Eisen. 2001. Microbial genes in the human genome: lateral transfer or gene loss? Science **292**:1903–1906.

Scholl, E. H., J. L. Thorne, J. P. McCarter, and D. M. Bird. 2003. Horizontally transferred genes in plant-parasitic nematodes: a high-throughput genomic approach. Genome Biol. **4**:R39.

Shimodaira, H., and M. Hasegawa. 1999. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. Mol. Biol. Evol. **16**:1114–1116.

Smant, G., J. Stokkermans, Y. T. Yan et al. (13 co-authors). 1998. Endogenous cellulases in animals: isolation of beta-1,4-endoglucanase genes from two species of plant-parasitic cyst nematodes. Proc. Natl. Acad. Sci. USA **95**:4906–4911.

Stanhope, M. J., A. Lupas, M. J. Italia, K. K. Koretke, C. Volker, and J. R. Brown. 2001. Phylogenetic analyses do not support horizontal gene transfers from bacteria to vertebrates. Nature **411**:940–944.

Steele, R. E., S. E. Hampson, N. A. Stover, D. F. Kibler, and H. R. Bode. 2004. Probable horizontal transfer of a gene between a protist and a cnidarian. Curr. Biol. **14**:R298–R299.

Steenbakkers, P. J. M., W. Ubhayasekera, H. Goossen, E. van Lierop, C. van der Drift, G. D. Vogels, S. L. Mowbray, and H. den Camp. 2002. An intron-containing glycoside hydrolase family 9 cellulase gene encodes the dominant 90 kDa component of the cellulosome of the anaerobic fungus *Piromyces* sp strain E2. Biochem. J. **365**:193–204.

Sugimura, M., H. Watanabe, N. Lo, and H. Saito. 2003. Purification, characterization, cDNA cloning and nucleotide sequencing of a cellulase from the yellow-spotted longicorn beetle, *Psacothea hilaris*. Eur. J. Biochem. **270**:3455–3460.

Suzuki, K., T. Ojima, and K. Nishita. 2003. Purification and cDNA cloning of a cellulase from abalone *Haliotis discus hannai*. Eur. J. Biochem. **270**:771–778.

Thompson, J. D., T. J. Gibson, F. Plewniak, F. Jeanmougin, and D. G. Higgins. 1997. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res. **25**:4876–4882.

Tokuda, G., N. Lo, H. Watanabe, G. Arakawa, T. Matsumoto, and H. Noda. 2004. Major alteration of the expression site of

endogenous cellulases in members of an apical termite lineage. Mol. Ecol. **13**:3219–3228.

Tokuda, G., N. Lo, H. Watanabe, M. Slaytor, T. Matsumoto, and H. Noda. 1999. Metazoan cellulase genes from termites: intron/exon structures and sites of expression. Biochim. Biophys. Acta **1447**:146–159.

Tomme P., R. A. J. Warren, and N. R. Gilkes. 1995. Cellulose hydrolysis by bacteria and fungi. Adv. Microb. Physiol. **37**:1–81.

Ujino-Ihara, T., Y. Taguchi, K. Yoshimura, and Y. Tsumura. 2003. Analysis of expressed sequence tags derived from developing seed and pollen cones of *Cryptomeria japonica*. Plant Biol. **5**:600–607.

Wang, J., M. Ding, Y. H. Li, Q. X. Chen, G. J. Xu, and F. K. Zhao. 2003. Isolation of a multi-functional endogenous cellulase gene from mollusc, *Ampullaria crossean*. Acta Biochim. Biophys. Sin. **35**:941–946.

Watanabe, H., H. Noda, G. Tokuda, and N. Lo. 1998. A cellulase gene of termite origin. Nature **394**:330–331.

Watanabe, H., and G. Tokuda. 2001. Animal cellulases. Cell Mol. Life Sci. **58**:1167–1178.

Xu, B. Z., J. C. Janson, and D. Sellos. 2001. Cloning and sequencing of a molluscan endo-beta-1,4-glucanase gene from the blue mussel, *Mytilus edulis*. Eur. J. Biochem. **268**: 3718–3727.

Yan, Y. T., G. Smant, J. Stokkermans, L. Qin, J. Helder, T. Baum, A. Schots, and E. Davis. 1998. Genomic organization of four beta-1,4-endoglucanase genes in plant-parasitic cyst nematodes and its evolutionary implications. Gene **220**: 61–70.

Yang, Z. 1997. PAML: a program for phylogenetic analysis by maximum likelihood. Comput. Appl. Biosci. **13**:555–556.

Yokoe, Y., and I. Yasumasu. 1964. The distribution of cellulase in invertebrates. Comp. Biochem. Physiol. **13**:323–338.

Zhu, X. D., G. Mahairas, M. Illies, R. A. Cameron, E. H. Davidson, and C. A. Ettensohn. 2001. A large-scale analysis of mRNAs expressed by primary mesenchyme cells of the sea urchin embryo. Development **128**:2615–2627.

Zhu, Z., T. Zheng, R. J. Homer, Y. K. Kim, N. Y. Chen, L. Cohn, Q. Hamid, and J. A. Elias. 2004. Acidic mammalian chitinase in asthmatic Th2 inflammation and IL-13 pathway activation. Science **304**:1678–1682.

Martin Embley, Associate Editor

Accepted January 31, 2005