J Mol Evol (2006) 63:801–814
DOI: 10.1007/s00239-006-0031-0

JOURNAL OF MOLECULAR EVOLUTION

The Frequency of Eubacterium-to-Eukaryote Lateral Gene Transfers Shows Significant Cross-Taxa Variation Within Amoebozoa

Russell F. Watkins,[1] Michael W. Gray[2]

[1] Centre for Molecular Medicine and Therapeutics, Child & Family Research Institute, and Department of Medical Genetics, University of British Columbia, Vancouver, British Columbia V5Z 4H4, Canada
[2] Department of Biochemistry and Molecular Biology, Dalhousie University, Room 8F-2, Tupper Medical Building, 5850 College Street, Halifax, Nova Scotia B3H 1X5, Canada

Received: 7 February 2006 / Accepted: 20 April 2006 [*Reviewing Editor:* Martin Kreitman]

Abstract. Single-celled bacterivorous eukaryotes offer excellent test cases for evaluation of the frequency of prey-to-predator lateral gene transfer (LGT). Here we use analysis of expressed sequence tag (EST) data sets to quantify the extent of LGT from eubacteria to two amoebae, *Acanthamoeba castellanii* and *Hartmannella vermiformis*. Stringent screening for LGT proceeded in several steps intended to enrich for authentic events while at the same time minimizing the incidence of false positives due to factors such as limitations in database coverage and ancient paralogy. The results were compared with data obtained when the same methodology was applied to EST libraries from a number of other eukaryotic taxa. Significant differences in the extent of apparent eubacterium-to-eukaryote LGT were found between taxa. Our results indicate that there may be substantial inter-taxon variation in the number of LGT events that become fixed even between amoebozoan species that have similar feeding modalities.

Key words: Lateral gene transfer — Horizontal gene transfer — Eukaryotes — Eubacteria — Genome evolution — Amoebozoa — Acanthamoeba — Hartmannella — Dictyostelium — Expressed sequence tags

Correspondence to: Russell F. Watkins; *email:* rwatkins@cmmt.ubc.ca

Introduction

Lateral gene transfer (LGT) appears to be relatively frequent within the eubacteria (Doolittle 1999), and a growing body of evidence (Andersson 2005) indicates that it plays some role in the evolution of the eukaryotic genome as well. There are now a number of well-documented examples of LGT from eubacteria into eukaryotes, including one case in which prokaryotic operon structure is apparently conserved in the transferred genes (Andersson and Roger 2002). Little doubt remains that the phenomenon is real and not merely the artifactual result of sampling bias or phylogenetic reconstruction. Moreover, cases of LGT in the opposite direction, from eukaryotes into prokaryotes, have been described (Koonin et al. 2001), a finding that implies the existence of horizontal routes of transmission of genetic information both within and between domains of life. Nonetheless, both the magnitude of LGT within the eukaryotes and its patterns of distribution remain unclear.

Broadly speaking, efforts to quantify the occurrence of LGT fall into three distinct categories. Searches by means of phylogenetic reconstruction across the widest possible taxonomic range for a particular gene or subset of genes (e.g., Keeling and Palmer 2001; Zardoya et al. 2002; Harper and Keeling 2004) have revealed candidate LGT events that affect multiple eukaryotic species, events that appear to indicate a relatively frequent occurrence of eubacterium-to-eukaryote and eukaryote-to-eukaryote LGT. Conversely, an alternate method involves searching through large numbers of sequences in, e.g., an EST library, in order to select genes whose phylogeny is discordant with the overall pattern of evolutionary descent for a particular species (Archibald et al. 2003). More recently, with the advent of full eukaryotic genome sequence data, comprehensive screening of entire genomes for LGT has become tractable (Sicheritz-Ponten and Andersson 2001).

The first attempt at full-genome screening for LGT events was for the human genome itself (Lander et al. 2001); however, artifacts plagued the results and the study and its conclusions engendered widespread criticism (Andersson et al. 2001; Stanhope et al. 2001; Salzberg et al. 2001). This human genome analysis was based solely on examination of E-values returned as the result of BLAST searches against GenBank, and it served to illustrate the shortcomings of an approach where similarity measures alone are used as a guide to uncovering LGT events. Nonetheless, with appropriate care to address potential sources of error, analysis of full-genome sequence data clearly represents a powerful strategy toward the goal of understanding the contribution of LGT to eukaryotic evolution. Such analyses are now being reported (e.g., Huang et al. 2004), but complete genome data will not be available any time soon for the full diversity of the eukaryotic radiation, constraining the application of this particular approach. For cross-taxa comparisons of rates and mechanisms of LGT, the more cost-effective generation and screening of data from sources such as EST libraries remains a necessity.

The full-genome sequences of *Dictyostelium discoideum* and *Entamoeba histolytica* have recently been published (Eichinger et al. 2005; Loftus et al. 2005), and in each case a genome-wide analysis was undertaken to survey the occurrence of genes that have apparently been laterally transferred from eubacteria into each species. *E. histolytica* was reported to contain 96 laterally transferred genes, which is an appreciable fraction of the total number of 9938 genes in this reduced genome. It may be that adaptation to the unique constraints of a parasitic, anaerobic lifestyle has selected for an increased frequency of fixation of laterally transferred DNA sequences, as many of the genes that appear to have been "adopted" in *E. histolytica* seem to be directly related to these functions. With the *D. discoideum* genome, a different methodology was used to estimate the horizontally transferred component, found to comprise only 18–22 of 13,541 protein-coding genes in total, a much smaller proportion than was found in *E. histolytica*. As in *E. histolytica*, these *D. discoideum* genes may confer direct adaptive advantages on the organism.

Recently Andersson (2005) has concluded that the results of a number of studies (e.g., Figge et al. 1999; Qian and Keeling 2001; Andersson et al. 2005) indicate that rates of LGT into phagotrophic eukaryotes are higher than those exhibited by nonphagotrophic taxa such as animals and fungi. Additionally it has been argued that the rate of occurrence of such LGT events varies significantly across eukaryotic lineages. Broad surveys of the incidence of LGT in specific taxa (Andersson and Roger 2003; Andersson et al. 2003) have suggested that LGT is a nontrivial source of genetic diversity in at least some portions of the eukaryotic radiation. Nonetheless, many questions remain regarding the extent to which the incidence of LGT varies between lineages and, more specifically, within individual lineages. It seems likely that many traits of an organism, such as trophic strategy, ploidy, reproductive mode, environmental complexity, and genomic lability, will interact to promote or discourage the incidence of LGT.

Through the Protist EST Program (PEP; http://megasun.bch.umontreal.ca/pepdb/pep.html), we are constructing and sequencing cDNA libraries for several amoebozoons, including *Acanthamoeba castellanii* and *Hartmannella vermiformis*. These two protists actively phagocytose prey bacteria and, in addition, are known to be capable of harboring a wide variety of microbes in intracellular associations (Horn and Wagner 2004), including important human pathogens (Kuiper et al. 2004). In at least one case (Jeon 2004) endosymbiotic eubacteria appear to show considerable biochemical complexity in their association with an amoebal host. Because *A. castellanii* and *H. vermiformis* exploit niches that involve ongoing and relatively elaborate interactions with eubacteria, they appear to be excellent candidates for the examination of LGT between eubacteria and eukaryotes. For this reason, our initial efforts at quantifying the rates and patterns of LGT within Amoebozoa, one of six recently described eukaryotic supergroups (Simpson and Roger 2004), have focused on these two amoebozoons.

In order to analyze roughly comparable data, we have selected similar-sized EST sets from other organisms within Amoebozoa, within opisthokonts (animals + fungi) broadly, and within several nonamoebozoan protist lineages. Our approach has been to utilize stringent criteria in the consideration of candidate cases of LGT. We perform multiple screening steps on EST clusters with the aim of removing all sequences for which various types of artifactual biases could reasonably explain the observed phylogenetic affiliation of the cluster. Accordingly, we retain only eukaryotic sequences that appear to show unambiguous affiliation with eubacteria, to the exclusion of the broad eukaryotic radiation. The results presented here provide an overview of the relative numbers of unique LGT events in several amoebozoan taxa in comparison

Table 1. Application of the screening algorithm to various protistan EST data sets

Organism	ESTs	Clusters	Contigs		Candidates	%	Screened Positive	%	Tree	%
Hartmannella[a]	11,686	5,205	1,234		16	1.30	12	0.97	9 (7, 6)	0.73 (0.57, 0.49)
					30	2.43	19	1.54	13 (10)	1.05 (0.81)
Acanthamoeba[a]	11,933	5,801	1,363	G	10	0.73	5	0.37	1 (1)	0.07 (0.07)
					25	1.83	13	0.95	3 (1)	0.22 (0.07)
*Dictyostelium*_VF	12,000	2,369	1,023	G	12	1.17	3	0.29	0	0.00
					20	1.96	7	0.68	0	0.00
*Dictyostelium*_AF	12,298	2,616	1,068	G	5	0.47	2	0.19	0	0.00
					9	0.84	5	0.47	3 (1, 1)	0.28 (0.09, 0.09)
Drosophila	12,000	4,894	1,645	G	8	0.49	4	0.24	2 (2)	0.12 (0.12)
					12	0.73	4	0.24	2 (2)	0.12 (0.12)
Toxoplasma	12,000	4,507	1,290	G	1	0.08	0	0.00	0	0.00
					3	0.23	0	0.00	0	0.00
Chlamydomonas[a]	12,000	9,811	1,040	G	3	0.29	0	0.00	0	0.00
					5	0.48	1	0.10	0	0.00[b]
Entamoeba[c]	—	—	9,968	G	—	—	—	—	96	0.97

Note. Screening was performed as described in the text, and the results of screening at a ΔE value of 30 (first line) or 20 (second line) are shown for each organism. The first set of numbers in parentheses refers to candidates remaining after removal of sequences where the direction of lateral transfer is uncertain, and the second set of numbers in parentheses refers, when it appears, to candidates remaining after further removal of sequences that likely do not represent amoebozoan lineage-specific LGT events. **G** indicates that genomic sequence data are available for some or all of the LGT candidate sequences. *Dictyostelium*_AF—the full EST data set for *D. discoideum* EST library from an aggregative-stage *Dictyostelium* (http://www.csm.biol.tsukuba.ac.jp/catalogue/Catalogue.html); *Dictyostelium*_VF—the first 12,000 ESTs from a vegetative-stage library for *D. discoideum* (http://www.csm.biol.tsukuba.ac.jp/catalogue/Catalogue.html); *Chlamydomonas*—the first 12,000 ESTs from a normalized *C. reinhardtii* EST library (no. 894; ftp://ftp.biology.duke.edu/pub/chlamy_genome/sequences/ESTclones/) made from pooled RNA of *C. reinhardtii* grown under a variety of conditions; *Toxoplasma*—12,000 random EST sequences from *T. gondii*, downloaded from dbEST and representing both tachyzoite and bradyzoite sequences; *Drosophila*—12,000 random *D. melanogaster* ESTs downloaded from dbEST.

[a] Clustering included base quality values in these taxa.

[b] The *Chlamydomonas* candidate was eliminated in the final (phylogenetic tree-based) screening step because it demonstrated strong affinity to Cyanobacteria.

[c] See Loftus et al. (2005a) for full-genome phylogenetic analysis.

with one another and with nonamoebozoan eukaryotes.

Materials and Methods

Selection of ESTs

A. castellanii (strain Neff) was cultured for 4 days at 30°C in Neff medium (Neff et al. 1964). *H. vermiformis* was grown for 14–16 days at room temperature in peptone/yeast extract/glucose medium (PYG, ATCC 712) modified by A. Lohan from pH 6.5 to pH 6.0–6.1. In both cases amoebae were harvested by centrifugation at 900*g* for 10 min and cell pellets were resuspended in 10 vol Trizol (Invitrogen), then flash-frozen in liquid nitrogen and stored frozen at −75°C. Library construction for *A. castellanii* and *H. vermiformis* was performed by DNA Technologies Inc. (Gaithersberg, MD, USA).

Raw sequence reads for *H. vermiformis* (11,686) and *A. castellanii* (11,933), along with base quality values as inferred by PHRED (Ewing et al. 1998), were processed by the CAP3 algorithm (Huang and Madan 1999) to create a list of clustered and singleton reads. For *H. vermiformis*, the initial data set of 11,686 ESTs yielded 1234 contigs containing two or more ESTs and 3971 singletons. For *A. castellanii* the original data set of 11,933 reads produced 1363 contigs and 4438 singletons. Singleton ESTs contain a higher proportion of problematic data, including lower-quality sequence, repetitive sequences, and short, improperly terminated cDNAs, than do clusters. Because we were particularly interested in the ratio of genes that might have originated via LGT vs. genes inherited in a strictly vertical fashion, it seemed prudent to remove the singleton reads from consideration, although this action simultaneously reduced the size of the data set. Allowing the singleton ESTs to remain could artificially deflate the ratio of LGT candidates to total examined clusters due to the inclusion of these noisy data. For this reason, the current study only examines trends in the more limited set of genes for which at least two ESTs are present in the original data.

For comparison, EST data sets from four taxonomically diverse eukaryotes were selected: two sets consisting of 12,000 and 12,298 ESTs, respectively, from vegetative (*Dicty*-VF) and aggregative (*Dicty*-AF) forms of *D. discoideum*, and one EST subset comprising 12,000 EST reads for each of *Drosophila melanogaster*, *Toxoplasma gondii* (phylum Apicomplexa), and *Chlamydomonas reinhardtii* (phylum Chlorophyta). All data sources are listed in the Note to Table 1. Each data set was clustered using CAP3, utilizing base quality values where available.

EST contigs were selected from the clustered data and used to search the GenBank nonredundant (nr) database, employing the BLASTALL software (BLASTX, default parameters [Altschul et al. 1997]). Proprietary PERL scripts along with data derived from the BLAST taxonomy database were used to separate the BLAST reports into taxonomic categories. Additional PERL scripts were then applied to screen the BLAST results both by sequence similarity to eubacterial and eukaryotic sequences and by precise taxonomic affiliation.

Screening for LGT Candidates

The criterion in primary screens for LGT candidacy was the degree of difference in apparent similarity to prokaryotic vs. eukaryotic sequences as measured by BLAST E-values (Fig. 1). This measure

- ● Primary Screening (GenBank)

  Search EST contigs against GenBank nr database by BLASTX
  - Ignoring hits against taxa from the same phylum, eliminate all contigs that are not more similar to eubacteria than eukaryotes by at least a specified difference in E-value exponents (ΔE).

- ● Secondary Screening (Additional Databases of Eukaryotic Diversity)

  Search remaining contigs against dbEST est_others, HTGS, GSS databases by BLASTX
  - Using the same criteria as in step 1, eliminate contigs that do not pass this screening step.

  Search remaining contigs against proprietary PEPdb database of eukaryotic EST contigs by BLASTX
  - Using the same criteria as in step 1, eliminate contigs that do not pass this screening step.

- ● Tertiary Screening (Phylogenetic Screening)

  For those contigs remaining, assemble all significant eukaryotic database hits, as well as representative prokaryotic database hits
  - Cluster all sequences, trim ambiguous regions.
  - Produce bootstrapped maximum likelihood trees for each candidate
  - Eliminate all candidates that cluster with other eukaryotic sequences and all candidates that can be linked to other eukaryotic sequences by means of one or more sub-50% nodes.

Fig. 1. Outline of the stepwise approach used in screening EST clusters.

can be problematic (see below) but, nevertheless, is a useful initial screening step. In two parallel analyses in this first-order screening, sequences were passed as LGT candidates if BLAST E-values to eubacterial sequences exceeded BLAST E-values to eukaryotic sequences by 20 and 30 orders of magnitude, respectively (hereinafter referred to as $\Delta E = 20$ and $\Delta E = 30$).

At the secondary stage, multiple additional databases (GenBank est_others, Genome Survey Sequence [GSS], High Throughput Genome Sequencing [HTGS]) of eukaryotic DNA sequences were screened manually for the presence of apparent orthologues to the LGT candidates. Exhaustive searches across these additional databases explored most of the publicly available eukaryotic DNA sequence data.

A critical additional step in secondary screening was a search against the proprietary Protist EST Program database (TBestDB; URL of the public version, PEPdbPub, is http://tbestdb.bcm.umontreal.ca/searches/login.php), which currently contains over 197,000 clusters from 67 taxa spanning the bulk of the eukaryotic radiation. TBestDB contains substantial sampling from otherwise unexamined eukaryotic phyla from all of the supergroups mentioned in a recent review (Simpson and Roger 2004), including Rhizaria, Excavata, Plantae, Chromalveolata, Amoebozoa, and Opisthokonta. All sequences in TBestDB are scheduled for eventual release to GenBank. Importantly, because TBestDB continued to expand in content during the course of this study, secondary screening was performed repeatedly to sample all relevant EST data, including those singleton ESTs that had been discarded initially. Thus, for all amoebozoons included in the present study, this screening step analyzed all available data, which, in the case of *A. castellanii*, continued to increase to nearly 20,000 EST reads.

Finally, candidate LGT clusters that exhibited affiliation with eubacterial taxa to greater than the established screening thresholds compared with eukaryotic taxa, and for which no clear eukaryotic orthologues could be identified, were passed through to phylogenetic analysis. In order to eliminate purely rate-based effects on BLAST similarity scores (Andersson et al. 2001), all potential eukaryotic orthologues and/or paralogues from database scans with E-values of e−05 or less were included in phylogenetic reconstructions. Amino acid sequences derived from the DNA sequences of the clusters were aligned with all eukaryotic hits from all databases. Prokaryotic sequences were selected from among all hits both to maximize taxonomic breadth and to select from among a wide range of significance levels in the respective matches. This selection was performed manually in each case, with choices including as broad a representation of prokaryotic taxa as possible while at the same time maintaining a tractable number of sequences for maximum likelihood-based analysis.

Phylogenetic Analysis of LGT Candidates

Alignments were generated using CLUSTALX (Chenna et al. 2003) with default alignment parameters and then edited manually with SEAVIEW (Galtier et al. 1996) to eliminate regions of questionable identity. PHYLIP-format files generated from these edited alignments were analyzed with TREE-PUZZLE 5.1 (Schmidt et al. 2002) to calculate an α parameter for Γ-corrected models of sequence evolution. Bootstrap replicates of the sequence data files were generated using SEQBOOT from the PHYLIP 3.6 package (Felsenstein 1989), and the resulting bootstrapped sequences were analyzed with PROML using the JTT + Γ model with eight rate categories.

A total of 50 bootstrapped maximum likelihood trees were constructed (included in Supplemental Materials) for sequences that passed the database screening steps. In each case the resulting consensus tree was inspected manually. Sequences that clustered strongly with additional eukaryotic phyla were excluded from LGT status, as were any sequences that could be clustered with additional eukaryotic phyla by means of nodes with bootstrap values $< 50\%$. The only sequences that were accepted as LGT candidates in the tertiary analysis were those that were restricted to their originating phylum alone, and which clustered specifically within Eubacteria at $> 50\%$ bootstrap support. Tallies of the numbers of LGT candidates per taxon were compared with one another by means of *Z* tests of proportionate differences.

Results

Primary screening of EST contigs initially showed a slightly higher number of LGT candidates in most of the amoebozoan taxa relative to the other eukaryotic taxa examined (Table 1). At the $\Delta E = 20$ screening level, the numbers of amoebozoan candidates ranged from barely 1% to > 2% of all clusters, whereas the other eukaryotes displayed a lower range, roughly between 0.2% and 0.7%. This range agrees with the original findings of Lander et al. (2001), in which roughly 1% of all human genes showed some evidence of a prokaryotic LGT origin based solely on similarity scores; however, the number is higher than the updated levels for the human data, revised after taking into account multiple sources of error (Salzburg et al. 2001). At the higher stringency screening level of $\Delta E = 30$, there is less apparent difference in LGT numbers between the amoebozoons and the other eukaryotic taxa.

The secondary screening step resulted in a dramatic drop in the number of LGT candidates in all taxa examined. This result is directly concordant with both the results and the recommendations of several papers that have addressed this issue previously. In the case of *T. gondii*, all LGT candidates were immediately eliminated from consideration at this stage. For *C. reinhardtii*, all but one of the candidate LGT events was eliminated, consistent with the results of Archibald et al. (2003) for this purely photosynthetic eukaryote. At the $\Delta E = 20$ screening level, approximately half of all amoebozoan candidate LGT events were found to have apparent eukaryotic orthologues, judged solely by this one criterion of similarity. In all cases in which these orthologues were subsequently examined by reconstruction of phylogenetic trees, clear orthology was confirmed. In some cases very limited taxonomic distributions of eukaryotic hits were found, and the possibility exists that these cases may represent LGT events that are in effect synapomorphies linking basal eukaryotic lineages. Nonetheless, they were eliminated from consideration according to the strict criteria of the present study.

Screening of the remaining candidate contigs by maximum likelihood analysis had nearly as dramatic an effect in detecting artifactual LGT results as did the secondary screening step itself (Table 1). Of the total of 50 eukaryotic contigs that, solely on the basis of database searches, showed no strong sequence similarity to any eukaryotic genes, 28 nonetheless clustered either clearly (> 50% bootstraps) or potentially (< 50% bootstraps) with other eukaryotic sequences and apart from prokaryotic sequences when rigorous phylogenetic reconstruction was applied. For most cases in which contigs were retained as LGT candidates, the reconstructed phylogenies were sufficiently robust that only the discovery of previously uncharacterized eukaryotic orthologues of the same genes is likely to dislodge them as putative LGT events from eubacteria into eukaryotes.

Two clusters from *H. vermiformis* were judged to be LGT candidates despite exhibiting very strong hits to eukaryotic ESTs in dbEST. In the case of *H. vermiformis* Contig 1030, a single EST from a *Pinus taeda* library (GenBank accession no. CF667347) matched the *H. vermiformis* sequence at a TBLASTX significance level of e−158. For Contig 664, a single EST from a *Sorghum bicolor* cDNA library (GenBank accession no. CD212320) was found at a TBLASTX significance level of e−117 relative to the *H. vermiformis* EST. Because in both cases the EST sequences also have extraordinarily high BLASTN levels of similarity to the *H. vermiformis* sequence (e−163 and 0.0, respectively), these sequences very likely represent contamination of the *Sorghum* and *Pinus* libraries with cDNAs from environmentally derived amoebozoons that were present in the tissue samples from which these libraries were created.

For *C. reinhardtii* we elected to remove from consideration one cluster that exhibited very strong bootstrap support for a cyanobacterial affiliation. With photosynthetic organisms it seems prudent to eliminate apparently cyanobacterial sequences from consideration, due to their possible origin from the protoplastid endosymbiont coupled with subsequent lineage-specific gene loss. However, in the case of *D. melanogaster*, we retained the two candidates for which a prokaryotic origin is strongly supported and that appear to be widely distributed across Metazoa. The specific characteristics of the phylogenetic trees reconstructed from these sequences are such that they could easily represent genes present in the eukaryotic ancestor but lost in multiple basal lineages. Lack of specific affiliation with prokaryotic taxa and basal branching relative to the prokaryotic radiation are two characteristics that have been cited among the diagnostic criteria for basal eukaryotic gene loss events rather than LGTs (Andersson et al. 2001). Nonetheless, because these sequences pass our specific criteria for LGT candidacy, we chose to retain them in this instance.

In neither case does the reversal of the above choices—i.e., retention of the *C. reinhardtii* candidate and removal of the *D. melanogaster* candidates—alter the significance of the *Z*-tests of proportional differences between taxa, as discussed below. A more difficult question to answer is whether one should remove from the analysis all candidates that show any α-proteobacterial affiliation. We elected not to do so, as proteobacteria appear to be frequent participants in LGT events, are very often found as endosymbionts in amoebozoons (Horn and Wagner

Table 2. Results of *Z*-tests of proportionate differences in numbers of LGT candidates in screened taxa

	Hart		*Acan*		*Dicty*-total		*Dicty*-A		*Dicty*-V		*Droso*		*Toxo*		*Chlamy*	
	20	30	20	30	20	30	20	30	20	30	20	30	20	30	20	30
A. *Acan*	**2.628**	**2.592**	—	—	—	—	—	—	—	—	—	—	—	—	—	—
B.	**2.775**	**2.185**														
C.	**2.775**	1.955														
A. *Dicty*-total	**2.819**	**3.011**	0.230	1.000	—	—	—	—	—	—	—	—	—	—	—	—
B.	**2.858**	**2.625**	0.133	1.000												
C.	**2.858**	**2.455**														
A. *Dicty*-A	**2.322**	**3.011**	0.295	1.000	—	—	—	—	—	—	—	—	—	—	—	—
B.	**2.637**	**2.653**	0.170	1.000												
C.	**2.637**	**2.455**														
A. *Dicty*-V	**3.625**	**3.011**	1.734	1.000	—	—	1.734	n/a	—	—	—	—	—	—	—	—
B.	**3.175**	**2.653**	1.000	1.000			1.000	n/a								
C.	**3.175**	**2.455**														
A. *Droso*	**3.075**	**2.365**	0.643	−0.427	0.446	1.415	0.869	−1.415	−1.415	−1.415	—	—	—	—	—	—
B.	**2.558**	1.934	−0.427	−0.427	−0.579	1.415	−0.220	−1.415	−1.415	−1.415						
C.	**2.558**	1.689														
A. *Toxo*	**3.625**	**3.011**	1.734	1.000	1.734	n/a	1.734	n/a	n/a	n/a	1.415	1.415	—	—	—	—
B.	**3.175**	**2.653**	1.000	1.000	1.000	n/a	1.000	n/a	n/a	n/a	1.415	1.415				
C.	**3.175**	**2.455**														
A. *Chlamy*	**3.625**	**3.011**	1.734	1.000	1.734	n/a	1.734	n/a	n/a	n/a	1.415	1.415	n/a	n/a	—	—
B.	**3.175**	**2.653**	1.000	1.000	1.000	n/a	1.000	n/a	n/a	n/a	1.415	1.415	n/a	n/a		
C.	**3.175**	**2.455**														

Note. Hart, Hartmannella vermiformis; Acan, Acanthamoeba castellanii; Dicty, Dictyostelium discoideum; Droso, Drosophila melanogaster; Toxo, Toxoplasma gondii; Chlamy, Chlamydomonas reinhardtii. *Z*-tests were also applied against a combined *D. discoideum* data set (denoted *Dicty*-total, with 1647 total clusters). Results for which the differences in LGT proportions in the data are significant at 95% are indicated in boldface. Test statistics were generated using the formula $Z = (p_1 - p_2)/\text{sqr}\ (p_1(1 - p_1)/n_1) + (p_2(1 - p_2)/n_2))$ (two-tailed, 95% critical value = 1.96). Row A shows test results with all LGT candidates included. Row B shows test results after exclusion of candidates for which the direction of LGT is uncertain. Row C shows test results with further exclusion of candidates for which an early amoebozoan LGT event is likely and, thus, for which the cluster is not reflective of lineage-specific LGT. Since this only affects the *H. vermiformis* tests, only these results are shown.

2004), and thus are excellent LGT donor candidates. Specific features of the phylogenetic tree that might indicate derivation of these sequences from the mitochondrial endosymbiont rather than LGT should be looked for carefully in these cases; in this regard, we can discern no such features.

Full-genome sequences are available for *D. discoideum* and *Drosophila melanogaster*, and a partial genome sequence is available for *A. castellanii* (Anderson et al. 2005). Using these resources, we found that all of the *D. discoideum* and *D. melanogaster* candidate LGT clusters are authentic, having clear genomic counterparts; in the case of *A. castellanii*, two of three candidate clusters are also represented within the sequenced portion of the genome. We can thus rule out spurious bacterial contamination of the cDNA libraries as a source of artifactual results in these instances. For the remainder of the candidates, further work will be necessary to demonstrate conclusively their genomic origins, although we do not regard bacterial contamination as a significant potential source of error.

Results of comparisons of the proportions of LGT candidates in each library are shown in Table 2. The salient feature is that the *H. vermiformis* results indicate a significantly greater extent of LGT in this species than in any of the other analyzed taxa, including the other amoebozoons. No LGT candidates were found at all in the *D. discoideum* VF library, so as an additional analysis we combined the AF and VF data sets (= *Dicty*-total in Table 2), for a total of 1647 nonredundant clusters with three unique LGT candidates. This operation, also, does not alter the significance of any of the comparisons.

Because some of the LGT candidates have an extremely limited phylogenetic distribution, the direction of LGT is unclear in these cases. Instances of eukaryote-to-prokaryote LGT are known (e.g., Schlieper et al. 2005), and it is possible that a gene found in a single eukaryote plus a very limited number of prokaryotes has been transferred in the eukaryote-to-prokaryote direction. Consequently we performed the tests after removal of all questionable candidates indicated by asterisks in Table 2 (B tests). In the majority of cases this operation had no effect on the significance of the tests. In the specific case of the *Hartmannella*-to-*Drosophila* comparison at the ΔE = 30 screening level, the test result falls just below the 95% significance level.

Table 3 shows the screening levels, annotations, and accession numbers for all of the candidate LGT events detected. No common theme appears to unite

Table 3. A list of all confirmed LGT candidates

Cluster no.			Screening level	Cluster identity	Accession no.
Hartmannella vermiformis					
863			30	Oxidoreductase (COG 667)	DQ384271
918	□		30	Unknown	DQ384272
465			30	Unknown	DQ384266
59			30	Aldehyde dehydrogenase (COG 1012)	DQ373922
664			30	Sugar transporter (COG 477)	DQ384270
277	□		30	Unknown	DQ384265
1030		◆	30	Glycogen debranching enzyme	DQ384273
480			30	Unknown (COG 1524)	DQ384267
102			30	Kinase (COG2187)	DQ384264
496			20	Aminopeptidase (COG 2234)	DQ386146
584			20	Aminopeptidase (COG 0308)	DQ384268
627			20	NADPH:quinone reductase (COG 604)	DQ384269
1091	□	◆	20	Hydrolase (COG 1073)	DQ384274
Dictyostelium discoideum AF					
548	□	◆	20	Adhesin AidA-like (COG 3468)	XP_636487
685			20	Permease (COG 477)	XP_635875
737	□	◆	20	Permease (COG 477)	AAO51569
Acanthamoeba castellanii					
452			30	Rhodanese-related sulfurtransferase (COG 2897)	DQ373920
1140	□		20	ABC-type phosphate/phosphonate transport system (COG 3221)	DQ373919
1203	□		20	Acetyltransferase (COG 3153)	DQ373921
Drosophila melanogaster					
645			30	Metallopeptidase	CAA65632
1462			30	Alkaline phosphatase (COG 1785)	NP_572742

Note. Library-specific cluster numbers, screening levels, cluster identities, and accession numbers are listed. Clusters removed in additional tests of proportionate differences due to unclear direction of LGT are indicated by □. Clusters removed in additional tests of proportionate differences due to clear evidence of their presence in multiple amoebozoan lineages (and hence that they are not lineage-specific LGT events) are indicated by ◆.

the functional relationships of these genes, although there might arguably be a general bias toward genes that participate in the mobilization and metabolism of environmental materials. Two of the *H. vermiformis* contigs (1030 and 1091) have limited numbers of clear orthologues among other amoebozoons, being found in *Physarum polycephalum* and *Mastigamoeba balamuthii* EST libraries, respectively. For *D. discoideum*, both the 548 and the 737 contigs also have other amoebozoan orthologues, contig 737, in particular, having a wide distribution within Amoebozoa. In the case of *D. melanogaster* each of the LGT candidates has a very wide-ranging metazoan distribution.

Since within the amoebozoan taxa we are specifically comparing lineage-specific amounts of LGT, our tests also include cases in which clusters are not considered due to their presence in multiple amoebozoan lineages (Table 2, C tests). In this case absence from any particular amoebozoan lineage is likely to represent either lineage-specific loss of a gene acquired in an early LGT event or a common ancestral event wherein the transferred gene is widely distributed within the clade in question but is effectively invisible elsewhere due to inadequate sampling. In fact this consideration only affects the results for *H. vermiformis* (Table 1), since for one of the two LGT candidates for this organism, the cluster has already been removed owing to a lack of clear directionality (Table 3). The two *D. discoideum* clusters that are found in other amoebozoan taxa are similarly already eliminated for the same reason, and thus the comparisons between groups are only affected for the *H. vermiformis* tests. In this case, for *H. vermiformis* against *A. castellanii*, the ΔE = 30 test falls just below the 95% significance level, but the levels of support do not change substantially for any other tests. As a result, only the ΔE = 20 test can be said to support a substantial difference between *A. castellanii* and *H. vermiformis*, but the general trend of a greater amount of LGT in *H. vermiformis* relative to other eukaryotes remains. A similar argument could be used to justify the removal of both candidate *D. melanogaster* events, but this operation does not result in any dramatic changes in the test results except that the *H. vermiformis* ΔE = 30 tests once again achieve significance. In any event, we are interested in the lineage-specific comparison against the entire phylum Metazoa.

Since whole-genome LGT analyses have now been reported for two amoebozoons, each of the 19 LGT candidates that passed tertiary screening in the present study was in turn screened against the two sets of published results. In no case did any of our amoebozoan LGT candidates appear to be substantially similar to any of the putative laterally transferred genes identified in *E. histolytica* by Loftus et al. (2005). Similarly, our amoebozoan LGT candidates (including the three from *D. discoideum*) did not match any of the 18 candidate laterally transferred genes identified by Eichinger et al. (2005) for *D. discoideum*. Two of the published candidate LGT events (ThyA and IPT) are in fact eliminated by the criteria of our EST screens. ThyA is found to have clear eukaryotic orthologues in the TBestDB database, information unavailable to Eichinger et al. (2005). IPT, also, is found to have multiple strong eukaryotic hits in TBestDB, a result that eliminates it under the criteria of our analysis. Two additional candidates, CnaB and an S13 peptidase, are found in the EST data sets for *D. discoideum* but are not reported by our algorithm. In these two cases, elimination of candidate sequences is a spurious result that is unavoidable when EST data are used for screening, a point discussed below.

Discussion

Methodological Considerations

In this study, we employ uncompromisingly rigorous criteria to scan for the incidence of strongly supported LGT events within each of the taxa examined. Our strategy is sufficiently stringent that we likely eliminate entirely from consideration certain potential classes of LGT: for instance, events in which orthologues of a given prokaryotic gene are laterally transferred into a variety of eukaryotic lineages, or LGT events involving highly conserved genes exhibiting pronounced sequence similarity across multiple domains of life. However, as a consequence of this stringency the LGT candidates we do retain are very strongly supported, and there is little doubt about their authenticity or concern about the role played by biases of phylogenetic reconstruction in their selection. These strongly supported cases of LGT can then be used to assess the relative amounts of one particular class of LGT in a variety of lineages, because these cases have been selected by consistently applied criteria.

The stringency of our screening criteria ensures that we will have removed as many false positives from our data set as is possible with current knowledge. Of course, without complete genome sequences from a very wide range of eukaryotic species, it is not possible to totally discount sparse-database artifacts as a source of error. However, as both the breadth and depth of sequence data increase, the likelihood of this particular artifact will decrease. The numbers of LGT candidates found at individual screening levels are maximal values using our particular criteria, rather than estimates of the absolute rate of LGT itself. The important aspect of these results is the relative numbers of candidate events per taxon. As discussed below, as the level of screening stringency is reduced, further LGT candidates are found, but their authenticity becomes increasingly questionable.

Our methodology will specifically exclude cases in which the same gene has been laterally transferred from numerous source taxa into different eukaryotic lineages. Such cases have been reported (Andersson et al., 2003), although some concerns may arise about the susceptibility of individual data sets to artifacts reflecting rate or composition bias. In the event that particular subsets of genes are laterally transferred with dramatically increased frequency relative to the bulk to the genome, we will underestimate the frequency of LGT to the same extent. On the other hand, if such cases exist at one end of a purely random distribution, then we would expect them to occur less often than the singular events. These are two distinct possibilities for the distribution of laterally transferred genes, which should produce demonstrably different results when the numbers of single-incidence events are compared to multiple-incidence ones.

It may well be that the selective advantage conferred by particular genes in particular circumstances (e.g., entry into an anoxic or low-oxygen niche) is strong enough to increase their representation in fixation events. Such cases do appear to exist in the literature, and since our methodology excludes them, we are in no position to consider their impact on the total amount of LGT. On the other hand, our current understanding of the constraints under which this process operates is not clear enough to warrant strong generalizations. Nevertheless, a comparison of the frequency of single-incidence LGT events across multiple taxa remains informative regarding the incidence of a particular class of LGT. If further analyses were to reveal that the incidence of multiple-order LGT events was significantly higher than single-order events, this would be an interesting and informative result.

The use of BLAST E-values in isolation to screen for LGT can be actively misleading (Andersson 2005) and this approach has correctly been criticized when used to search for laterally transferred genes (Katz 2002). While the E-value is specifically not a measure of phylogenetic relationship, the BLAST scores do provide a statistically sound measure of sequence similarity that can serve as a useful starting point in attempting to identify LGT candidates, providing careful and thorough phylogenetic screening is then applied (Hall et al. 2005). This principle represents a fundamental distinction between our methodology and those that rely on mass screening by phylogenetic analysis. Purely phylogenetic screening is usually based on construction of phylogenetic trees derived from automated alignments, resulting in a candidate set against which manual screening of the trees is then performed. This method will obviously retain candidates that our approach discards, but in our analysis we were particularly interested in eliminating poorly supported or potentially artifactual trees, retaining only the most convincing candidates. For this reason we are more comfortable with an approach that selects for candidates that are discernibly more eubacterial than eukaryotic in origin. A careful comparative analysis of the advantages of these two methods is probably called for, but our approach very likely selects for a strongly supported subset of the total group of candidates that might be found by mass phylogenetic screening.

On the other hand, there is a clear drawback to this methodology, in that the rate of false positives is decreased at the expense of a likely increase in false negatives. Since, as discussed below for *D. discoideum*, our methodology is able to provide nonoverlapping sets of LGT candidate genes even compared with the output of full-genome scans, it is actually rather difficult to estimate the total error rate. We judge that in this case a consistent standard applied across multiple taxa still serves as a useful estimator of the relative amounts of LGT that may have occurred; nevertheless, it must be clearly understood that this is an underestimate. Ours is an inherently conservative estimator of the total amount of LGT, but the resulting candidate set is likely highly enriched for nonartifactual events.

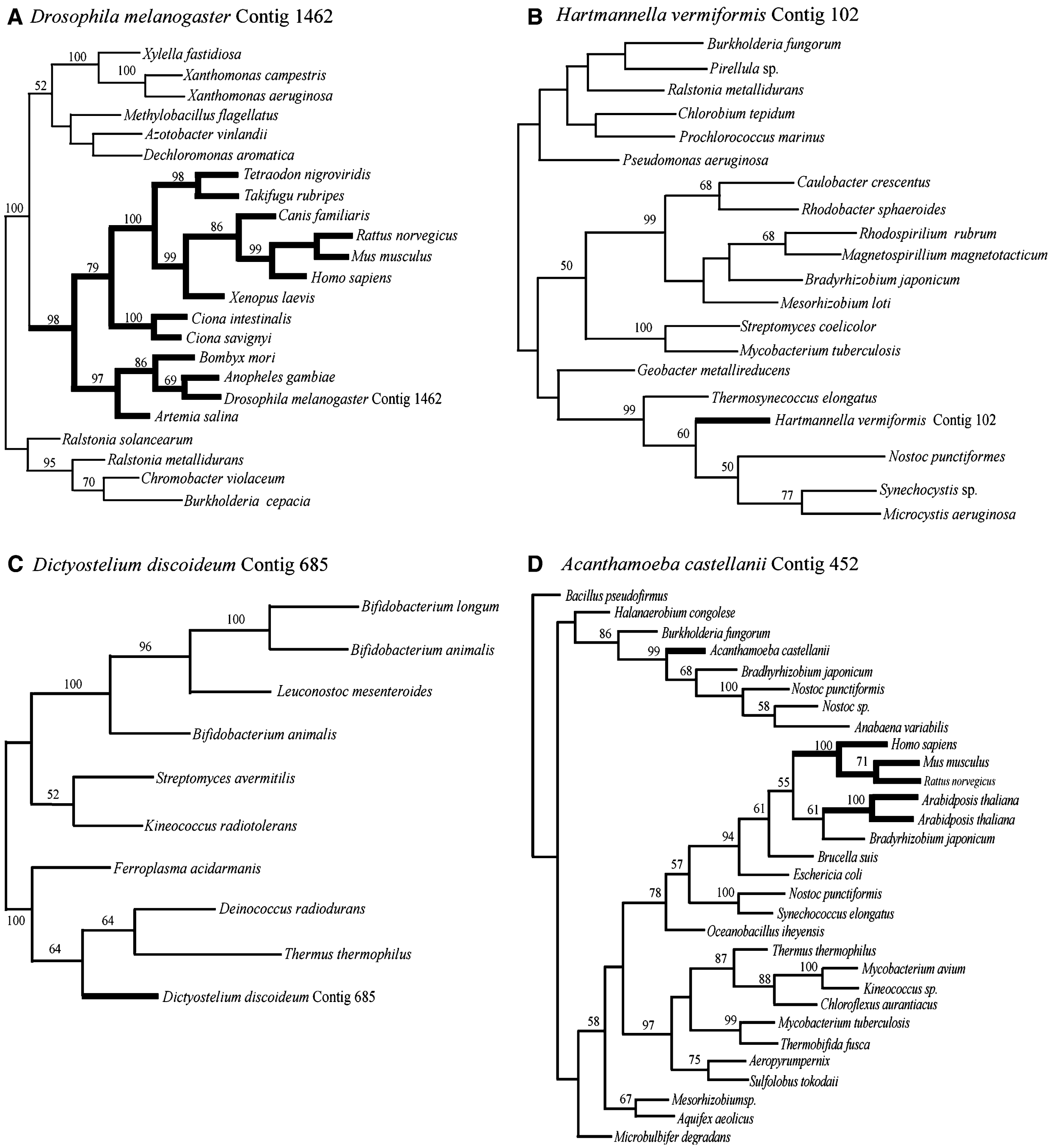
The most frequent cause for elimination of individual sequences from the candidate pool is the discovery of apparent eukaryotic orthologues through screening of additional databases. Typically half or more of all candidate LGT events initially selected are eliminated immediately by searches against additional databases of eukaryotic sequences. Sparse-database artifacts can be a major contributing factor to artifactual LGT candidacy (Salzburg et al. 2001) and our results reaffirm the vital importance of maximal screening for this effect in any examination of LGT. Lack of database completeness in both eukaryotic breadth (sampling across the widest possible range of eukaryotic taxa) and depth (proportion of the expressed genome represented in the database per organism) can be expected to contribute to this problem individually. Screening against the PEP database allowed us to minimize sampling artifacts to the greatest extent possible with currently available eukaryotic data. Phylogenetic screening of the remaining LGT candidates commonly eliminates a significant number of them. In every case this elimination is due to the clustering of the candidate LGT sequence with eukaryotic orthologues that appeared to be distant paralogues when judged solely by sequence similarity. In many cases nearly half of all remaining candidate LGTs were eliminated upon phylogenetic screening, a result that emphasizes the multiple sources of error that can plague the process of screening for these events.

Comparison with Nonamoebozoan Taxa

Our results with *C. reinhardtii* are entirely consistent with those of Archibald et al. (2003), who were unable to identify any LGT candidates among a subset of plastid-targeted genes in this organism. This finding contrasts sharply with the apparently higher rate of LGT that these authors inferred for the chlorarachniophyte alga *Bigelowiella natans*, which is known to phagocytose prey organisms in addition to being actively photosynthetic. This difference was interpreted as support for the ratchet model (“you-are-what-you-eat”) of Doolittle (1998).

When the results of our screening methodology as applied to *T. gondii* are compared with whole-genome analysis of LGT in other alveolates (Huang et al. 2004), it is apparent that our approach may exclude from consideration some candidates that are found when purely phylogenetic screening methods are applied. In the alveolate *Cryptosporidium parvum*, 31 genes of 5519 in the entire genome (around 0.56%) are found to cluster strongly with eubacterial taxa in preference to eukaryotic taxa. Several of these candidates have apparent orthologues in our *T. gondii* data set but are not picked out as LGT candidates by our screening algorithm. Many of these genes are included in our candidate set by relaxing the stringency criteria to accept very close matches to eukaryotic sequences. Others may be missed due to the effects of screening partial gene sequences derived from EST data, as discussed below.

Upon examining these cases in detail, we find that when additional eukaryotic data from TBestDB are included, reconstruction of phylogenetic trees for these candidates does usually show support for the affiliations reported by Huang et al. (2004). These groupings typically involve very highly conserved genes for which there is little strong differentiation between prokaryotic and eukaryotic sequences. In some cases there is apparent basal branching of the alveolate lineages relative to the bulk of the eukaryotic radiation, which raises serious concerns about long-branch effects. To what extent these clusterings are systematically affected by rate and other biases is unclear. Our methodology has the advantage of selecting for cases in which the effect of this particular artifact is minimized. Although purely phylogenetic



A *Drosophila melanogaster* Contig 1462

B *Hartmannella vermiformis* Contig 102

C *Dictyostelium discoideum* Contig 685

D *Acanthamoeba castellanii* Contig 452

Fig. 2. Phylogenetic trees for selected LGT candidates that passed all screening steps. Bootstrap replicates (100) were generated for each aligned and trimmed data set using the JTT + Γ model, with α calculated from the initial alignments by TREE-PUZZLE 5.1. **A** *D. melanogaster* alkaline phosphatase cluster; α = 1.19. **B** *H. vermiformis* kinase cluster; α = 1.47. **C** *D. discoideum* unidentified cluster; α = 3.27. **D** *A. castellanii* sulfur transferase-like cluster; α = 2.17.

screening does have some clear advantages over the methods reported here, great care must be taken in interpreting the results of the trees generated in this way because the screening method necessarily enriches for results in which biases of phylogenetic reconstruction are a factor.

The results for *D. melanogaster* make sense in light of the expected low rate of LGT in taxa with substantial germline/soma separation. The results we report here are in fact concordant with the levels reported by Salzberg et al. (2001) for apparent LGT into the basal metazoan radiation. Nonetheless, as discussed above, the specific characteristics of the trees generated for these clusters (Fig. 2)—robust basal branching combined with lack of affiliation to specific eubacterial lineages—make it likely that these candidates are spurious, the result of gene loss in basal eukaryotic lineages. These candidates are retained here on purely technical grounds in that they meet the specified screening criteria.

Comparison Within Amoebozoa

The apparent rate of eubacterium-to-eukaryote LGT in *D. discoideum* is also quite low. In comparing our results to the full-genome phylogenetic analysis for *D. discoideum* (Eichinger et al. 2005), the overall trends in the data are similar, in that there is no support for high levels of LGT. On the other hand, it is interesting that our analysis selects a different subset of LGT candidates than the analysis by Eichinger et al. (2005), which was based on a primary screening step for protein sequences that contain exclusively prokaryotic PFAM domains. Since all of the candidates that pass our analysis have fairly limited taxonomic distributions, they do not contain well-defined PFAM domains and thus do not show up in scans for these domains. On the other hand, our methodology discards some clusters that represent well-supported LGT events according to the PFAM screening methodology; however, this result is an unavoidable outcome of the use of EST clusters. Because cluster sequences are very frequently shorter than the full coding sequence, BLAST E-values will be reduced accordingly. For example, lowering our screening limit results in inclusion of the S13 peptidase candidate from the full-genome data set. Conversely, the unusually long CnaB gene, for which the pattern of BLAST hits against several apparently orthologous clusters is complex, remains excluded; this result can be attributed to the relatively short lengths of the available CnaB EST clusters. In the absence of full-genome data, the latter bias could be eliminated by obtaining full-length sequence data for all candidate genes prior to subsequent bioinformatics-based analysis, including primary screens against GenBank.

The number of LGT candidates in *H. vermiformis* is significantly greater than in the other taxa examined here, an observation that may also be consistent with Doolittle's (1998) ratchet model, wherein unicellular phagocytic eukaryotes would be expected to show a strong LGT bias due to ongoing cytoplasmic exposure to prokaryotic DNA. *A. castellanii*, on the other hand, exhibits a degree of apparent LGT that is statistically indistinguishable from the lower values observed at these same stringency levels for *D. discoideum* and the other eukaryotes in our study. The *A. castellanii* and *H. vermiformis* results appear to be robustly different at the $\Delta E = 20$ screening level, and difficult to reconcile. It is possible that *A. castellanii* inherently has a somewhat lower rate of LGT due to specific differences in gene regulation, phagocytic mechanisms, or other factors. On the other hand, the E-value criterion used for initial screening is likely itself sensitive to overall rate effects in individual organisms, so that closer examination of candidates at a wider range of stringencies may be revealing.

For a comparison of rates across taxa to be meaningful, the compared numbers must be estimators of the true numbers of unique lineage-specific fixations of laterally transferred genes. Inclusion of genes that have been acquired ancestrally in a given branch (for instance, in the common ancestor of *H. vermiformis* and *A. castellanii*) and then lost in one of these branches will be misleading. Some of these candidates can be removed from consideration by eliminating genes that are found in other amoebozoons, in which case a strong argument can be made that their acquisition predates the divergence of the two taxa in question. In the case of *H. vermiformis* contig 1030, this consideration is especially relevant given that a clear orthologue is present in *P. polycephalum*; thus, this gene may have been acquired very early in the amoebozoan radiation. When a search for such possibilities is carried out, our conclusions are not dramatically altered although the distinction between *H. vermiformis* and *A. castellanii* does become nonsignificant at the most stringent screening level (Table 2, C tests).

A more complex issue with sampling is that as more closely related taxa are examined, the number of available comparative data tends to become smaller. A candidate gene found in one species could be missed in the second species purely due to sampling effects. In the *H. vermiformis*-*A. castellanii* comparison, we are fortunate in having a partial (0.5X) genome sequence for *A. castellanii*. We can thus say with confidence that none of the *H. vermiformis* candidate genes is found in the sampled portion of the *A. castellanii* genome. Hence, in order for these candidates to represent Lobosa-specific genes, all of the LGTs that are currently classified as *H. vermiformis*-specific would have to be contained within the remaining unsampled portion of the *A. castellanii* genome. Additionally there do not appear to be any unusual numbers of LGT candidates in *A. castellanii* that are eliminated specifically upon consideration of the *H. vermiformis* data. In the absence of full-genome data from all taxa involved, these comparisons must remain tentative, particularly for taxa that are phylogenetically very close to one another.

In the case of any individual amoebozoan taxon compared against taxa from other phyla, the assumption of lineage specificity has already been tested to the greatest extent currently possible in searches against multiple databases. Since these searches include the complete proteomes of several eukaryotes and the best possible sampling of additional eukaryotic diversity, we can assert with confidence that these genes appear to be unique to individual lineages of Amoebozoa. As the amount of data and number of taxa available for comparison expands, an appeal to lineage-specific gene loss as an alternate explanation becomes less and less tenable.

Comparisons between taxa within a clade probably cannot be judged by the same standards as interclade comparisons. Since species within a clade (e.g., *A. castellanii* and *H. vermiformis*) have diverged more recently than have the clades themselves, the absolute distances between groups within the clade, in terms of both genetic distance and number of generations, are likely to be shorter. This may be problematic when the results of the intraclade tests are compared to those of the interclade tests. One approach to minimizing this potential artifact is to consider the results of tests for which apparently homoplastic LGT events have been removed in intraclade comparisons (the C tests in Table 2, which are only relevant for *H. vermiformis*), while considering the full set of LGT events for interclade comparisons (the B tests in Table 2). For interclade comparisons this approach has the effect of considering all events along each branch within each clade, while for the intraclade comparisons we are only considering events that appear to be lineage-specific. In practical terms this procedure does not alter substantially the interpretation of our results.

Because the RNA for both *A. castellanii* and *H. vermiformis* was extracted in a similar manner and library construction was performed with identical methodology, it seems unlikely that trivial explanations such as biases in RNA purification or library construction would lead to differences in the expression profiles evident for the two libraries. On the other hand, the organisms themselves are cultured rather differently: *A. castellanii* in a shaken liquid culture at 30°C; *H. vermiformis*, at room temperature in a purely stationary culture. Whether this difference in growth conditions alone promotes substantial differences in gene expression between the two organisms is unknown. A related issue is whether or not the occurrence of LGT varies across expression levels, the pattern of which will of course vary between lineages. It does seem possible that highly expressed genes are more likely to participate in recombinant events with exogenous DNA due to a less condensed chromatin state. For the moment (Table 3) there is no obvious indication that clusters that include larger numbers of ESTs are more frequently represented among the LGT candidates, but closer examination of this issue is probably warranted.

It is worth noting that reducing the screening threshold to $\Delta E = 5$ dramatically increases the number of primary candidate sequences for both *A. castellanii* and *H. vermiformis* (to 86 and 98, respectively). In contrast, the numbers for *T. gondii* and *C. reinhardtii* increase to only 15 and 16 candidates, respectively, whereas the *D. melanogaster* number increases to 47. On the other hand, the *D. discoideum* VF and AF libraries show an increase to only 54 and 30 candidates, respectively. It seems likely that the total number of good LGT candidates will increase when full phylogenetic screening is applied to the lower-stringency candidate sets from the two unicellular amoebae.

With the advent of full-genome sequences for *D. discoideum* and *E. histolytica*, as well as partial genome sequences for multiple additional *Entamoeba* species and for *A. castellanii*, full-genome comparisons of LGT within Amoebozoa are becoming tractable. The first two reports of full-genome analyses from *E. histolytica* and *D. discoideum* highlighted apparent differences between the two taxa: specifically, *E. histolytica* seems to have been affected by LGT to a greater extent than has *D. discoideum*. While the study reported here has added some additional LGT candidates for *D. discoideum* while at the same time likely eliminating at least one of the previously reported candidates, the broad pattern of our results is entirely compatible with the full-genome analysis. In the work reported here the amount of eubacterium-to-eukaryote LGT does not dramatically exceed 1% of all genes for any amoebozoon. While this is certainly a minimum number for *H. vermiformis* and *A. castellanii*, it nonetheless stands in sharp contrast to the very high levels of prokaryote-to-prokaryote LGT that have been reported (Doolittle et al. 2002). It is also important to note that despite our highly conservative methodology, which is bound to exhibit an increased rate of false rejections, our estimated values are actually very similar to those for *D. discoideum* and *E. histolytica* when full-genome sets are examined by alternate methods that do not specifically exclude multiple-event transfers. For *D. discoideum* we may have found some candidate laterally transferred genes that were not identified in the original full-genome study, but it seems unlikely that the number of candidates will grow substantially. In fact, the LGT number is much more likely to decrease as additional eukaryotic data are considered. The numbers of candidate events in both *A. castellanii* and *H. vermiformis* could, however, still increase as other methods of screening are applied. Nonetheless, the results of the present study, compared with those of the full-genome eukaryotic analyses published to date, suggest that LGT, while not an insignificant force in the evolution of eukaryotes, plays overall a relatively smaller role than in prokaryotes. The argument has been made (Kurland 2005) that LGT in eukaryotes may be limited by a number of inherent barriers. Nonetheless, the rate of eukaryote-to-eukaryote LGT remains to be carefully examined, and indeed there are some suggestions (Archibald et al. 2003; Keeling and Inagaki 2003) that it may represent a nontrivial component of the total amount of LGT in selected eukaryotes.

Acknowledgments. We thank C. Nesbø for her careful reading of the original draft of the manuscript and for her insightful comments. We also thank A. Roger for very useful discussions. This study was conducted under the auspices of the Protist EST Program (PEP), a Genome Canada large-scale genomics project, with funding through Genome Atlantic and the Atlantic Innovation Fund. M.W.G. is pleased to acknowledge salary support from the Canada Research Chairs Program and the Canadian Institute for Advanced Research (Program in Evolutionary Biology).

References

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402

Anderson IJ, Watkins RF, Samuelson J, Spencer DF, Majoros WH, Gray MW, Loftus BJ (2005) Gene discovery in the *Acanthamoeba castellanii* genome. Protist 156:203–214

Andersson JO (2005) Lateral gene transfer in eukaryotes. Cell Mol Life Sci 62:1182–1197

Andersson JO, Roger AJ (2002) Evolutionary analyses of the small subunit of glutamate synthase: gene order conservation, gene fusions, and prokaryote-to-eukaryote lateral gene transfers. Eukaryot Cell 1:304–310

Andersson JO, Roger AJ (2003) Evolution of glutamate dehydrogenase genes: evidence for lateral gene transfer within and between prokaryotes and eukaryotes. BMC Evol Biol 3:14

Andersson JO, Doolittle WF, Nesbø CL (2001) Are there bugs in our genome? Science 292:1848–1850

Andersson JO, Sjögren ÅM, Davis LAM, Embley TM, Roger AJ (2003) Phylogenetic analyses of diplomonad genes reveal frequent lateral gene transfers affecting eukaryotes. Curr Biol 13:94–104

Andersson JO, Sarchfield SW, Roger AJ (2005) Gene transfers from Nanoarchaeota to an ancestor of diplomonads and parabasalids. Mol Biol Evol 22:85–90

Archibald JM, Rogers MB, Toop M, Ishida K, Keeling PJ (2003) Lateral gene transfer and the evolution of plastid-targeted proteins in the secondary plastid-containing alga *Bigelowiella natans*. Proc Natl Acad Sci USA 100:7678–7683

Chenna R, Sugawara H, Koike T, Lopez R, Gibson TJ, Higgins DG, Thompson JD (2003) Multiple sequence alignment with the Clustal series of programs. Nucleic Acids Res 31:3497–3500

Doolittle WF (1998) You are what you eat: a gene transfer ratchet could account for bacterial genes in eukaryotic nuclear genomes. Trends Genet 14:307–311

Doolittle WF (1999) Lateral genomics. Trends Cell Biol 1999:M5–M8

Doolittle WF, Boucher Y, Nesbø CL, Douady CJ, Andersson JO, Roger AJ (2002) How big is the iceberg of which organellar genes in nuclear genomes are but the tip? Philos Trans R Soc Lond B Biol Sci 358:39–58

Eichinger L, Pachebat JA, Glöckner G, Rajandream M-A, Sucgang R, Berriman M, Song J, Olsen R, Szafranski K, Xu Q, Tunggal B, Kummerfeld S, Madera M, Konfortov BA, Rivero F, Bankier AT, Lehmann R, Hamlin N, Davies R, Gaudet P, Fey P, Pilcher K, Chen G, Saunders D, Sodergren E, Davis P, Kerhornou A, Nie X, Hall N, Anjard C, Hemphill L, Bason N, Farbrother P, Desany B, Just E, Morio T, Rost R, Churcher C, Cooper J, Haydock S, van Driessche N , Cronin A, Goodhead I, Muzny D, Mourier T, Pain A, Lu M, Harper D, Lindsay R, Hauser H, James K, Quiles M, Madan Babu M, Saito T, Buchrieser C, Wardroper A, Felder M, Thangavelu M, Johnson D, Knights A, Loulseged H, Mungall K, Oliver K, Price C, Quail MA, Urushihara H, Hernandez J, Rabbinowitsch E, Steffen D, Sanders M, Ma J, Kohara Y, Sharp S, Simmonds M, Spiegler S, Tivey A, Sugano S, White B, Walker D, Woodward J, Winckler T, Tanaka Y, Shaulsky G, Schleicher M, Weinstock G, Rosenthal A, Cox EC, Chisholm RL, Gibbs R, Loomis WF, Platzer M, Kay RR, Williams J, Dear PH, Noegel AA, Barrell B, Kuspa A (2005) The genome of the social amoeba *Dictyostelium discoideum*. Nature 435:43–57

Ewing B, Hillier L, Wendl MC, Green P (1998) Base-calling of automated sequencer traces using *phred*. I. Accuracy assessment. Genome Res 8:175–178

Felsenstein J (1989) PHYLIP—Phylogeny Inference Package (version 3.2). Cladistics 5:164–166

Figge RM, Schubert M, Brinkmann H, Cerff R (1999) Glyceraldehyde-3-phosphate dehydrogenase gene diversity in eubacteria and eukaryotes: evidence for intra- and inter-kingdom gene transfer. Mol Biol Evol 16:429–440

Galtier N, Gouy M, Gautier C (1996) SEAVIEW and PHYLO_WIN: two graphic tools for sequence alignment and molecular phylogeny. Comput Appl Biosci 12:543–548

Hall C, Brachat S, Dietrich FS (2005) Contribution of horizontal gene transfer to the evolution of *Saccharomyces cerevisiae*. Eukaryot Cell 4:1102–1115

Harper JT, Keeling PJ (2004) Lateral gene transfer and the complex distribution of insertions in eukaryotic enolase. Gene 340:227–235

Horn M, Wagner M (2004) Bacterial endosymbionts of free-living amoebae. J Euk Microbiol 51:509–514

Huang X, Madan A (1999) CAP3: a DNA sequence assembly program. Genome Res 9:868–877

Huang J, Mullapudi N, Lancto CA, Scott M, Abrahamsen MS, Kissinger JC (2004) Phylogenomic evidence supports past endosymbiosis, intracellular and horizontal gene transfer in *Cryptosporidium parvum*. Genome Biol 5:R88

Jeon KW (2004) Genetic and physiological interactions in the amoeba-bacteria symbiosis. J Eukaryot Microbiol 51: 502–508

Katz LA (2002) Lateral gene transfers and the evolution of eukaryotes: theories and data. Int J Syst Evol Microbiol 52:1893–1900

Keeling PJ, Inagaki Y (2004) A class of eukaryotic GTPase with a punctate distribution suggesting multiple functional replacements of translation elongation factor 1α. Proc Natl Acad Sci USA 101:15380–15385

Keeling PJ, Palmer JD (2001) Lateral transfer at the gene and subgenic levels in the evolution of eukaryotic enolase. Proc Natl Acad Sci USA 98:10745–10750

Koonin EV, Makarova KS, Aravind L (2001) Horizontal gene transfer in prokaryotes: quantification and classification. Annu Rev Microbiol 55:709–742

Kuiper MW, Wullings BA, Akkermans AD, Beumer RR, van der Kooij D (2004) Intracellular proliferation of *Legionella pneumophila* in *Hartmannella vermiformis* in aquatic biofilms grown on plasticized polyvinyl chloride. Appl Environ Microbiol 70:6826–6833

Kurland CG (2005) What tangled web: barriers to rampant horizontal gene transfer. Bioessays 27:741–747

Lander ES, Linton LM, Birren B, et al. (2001) Initial sequencing and analysis of the human genome. Nature 409:860–921

Loftus B, Anderson I, Davies R, Alsmark UCM, Samuelson J, Amedeo P, Roncaglia P, Berriman M, Hirt RP, Mann BJ, Nozaki T, Suh B, Pop M, Duchene M, Ackers J, Tannich E, Leippe M, Hofer M, Bruchhaus I, Willhoeft U, Bhattacharya A, Chillingworth T, Churcher C, Hance Z, Harris B, Harris D, Jagels K, Moule S, Mungall K, Ormond D, Squares R, Whitehead S, Quail MA, Rabbinowitsch E, Norbertczak H, Price C, Wang Z, Guillén N, Gilchrist C, Stroup SE, Bhattacharya S, Lohia A, Foster PG, Sicheritz-Ponten T, Weber C, Singh U, Mukherjee C, El-Sayed NM, Petri WA Jr, Clark CG, Embley TM, Barrell B, Fraser CM, Hall N (2005) The genome of the protist parasite *Entamoeba histolytica*. Nature 433:865–868

Neff RJ, Ray SA, Benton WF, Wilborn M (1964) Induction of synchronous encystment (differentiation) in *Acanthamoeba* sp. In: Prescott DM (ed) Methods in cell physiology, Vol I. Academic Press, New York, pp 55–83

Qian Q, Keeling PJ (2001) Diplonemid glyceraldehyde-3-phosphate dehydrogenase (GAPDH) and prokaryote-to-eukaryote lateral gene transfer. Protist 152:193–201

Salzberg SL, White O, Peterson J, Eisen JA (2001) Microbial genes in the human genome: lateral transfer or gene loss? Science 292:1903–1906

Schlieper D, Oliva MA, Andreu JM, Löwe J (2005) Structure of bacterial tubulin BtubA/B: evidence for horizontal gene transfer. Proc Natl Acad Sci USA 102:9170–9175

Schmidt HA, Strimmer K, Vingron M, von Haeseler A (2002) TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics 18:502–504

Sicheritz-Pontén T, Andersson SGE (2001) A phylogenomic approach to microbial evolution. Nucleic Acids Res 29:545–552

Simpson AG, Roger AJ (2004) The real 'kingdoms' of eukaryotes. Curr Biol 14:R693–R696

Stanhope MJ, Lupas A, Italia MJ, Koretke KK, Volker C, Brown JR (2001) Phylogenetic analyses do not support horizontal gene transfers from bacteria to vertebrates. Nature 411:940–944

Zardoya R, Ding X, Kitagawa Y, Chrispeels MJ (2002) Origin of plant glycerol transporters by horizontal gene transfer and functional recruitment. Proc Natl Acad Sci USA 99:14893–14896