SMITH COLLEGE

Smith ScholarWorks

Biological Sciences: Faculty Publications

Biological Sciences

8-31-2015

Recent Events Dominate Interdomain Lateral Gene Transfers Between Prokaryotes and Eukaryotes and, with the Exception of Endosymbiotic Gene Transfers, Few Ancient Transfer Events Persist

Laura A. Katz
Smith College, lkatz@smith.edu

Follow this and additional works at: https://scholarworks.smith.edu/bio_facpubs

 Part of the Biology Commons

Recommended Citation

Katz, Laura A., "Recent Events Dominate Interdomain Lateral Gene Transfers Between Prokaryotes and Eukaryotes and, with the Exception of Endosymbiotic Gene Transfers, Few Ancient Transfer Events Persist" (2015). Biological Sciences: Faculty Publications, Smith College, Northampton, MA.
https://scholarworks.smith.edu/bio_facpubs/98

This Article has been accepted for inclusion in Biological Sciences: Faculty Publications by an authorized administrator of Smith ScholarWorks. For more information, please contact scholarworks@smith.edu

PHILOSOPHICAL TRANSACTIONS B

rstb.royalsocietypublishing.org

Research

Cite this article: Katz LA. 2015 Recent events dominate interdomain lateral gene transfers between prokaryotes and eukaryotes and, with the exception of endosymbiotic gene transfers, few ancient transfer events persist. *Phil. Trans. R. Soc. B* **370**: 20140324.
http://dx.doi.org/10.1098/rstb.2014.0324

Accepted: 8 July 2015

One contribution of 17 to a theme issue 'Eukaryotic origins: progress and challenges'.

Subject Areas:
evolution, bioinformatics, microbiology

Keywords:
phylogenomics, endosymbiotic gene transfer, horizontal gene transfer, eukaryotic tree of life

Author for correspondence:
Laura A. Katz
e-mail: lkatz@smith.edu

Electronic supplementary material is available at http://dx.doi.org/10.1098/rstb.2014.0324 or via http://rstb.royalsocietypublishing.org.

THE ROYAL SOCIETY PUBLISHING

Recent events dominate interdomain lateral gene transfers between prokaryotes and eukaryotes and, with the exception of endosymbiotic gene transfers, few ancient transfer events persist

Laura A. Katz[1,2]

[1]Department of Biological Sciences, Smith College, Northampton, MA 01063, USA
[2]Program in Organismic and Evolutionary Biology, UMass-Amherst, Amherst, MA 01003, USA

While there is compelling evidence for the impact of endosymbiotic gene transfer (EGT; transfer from either mitochondrion or chloroplast to the nucleus) on genome evolution in eukaryotes, the role of interdomain transfer from bacteria and/or archaea (i.e. prokaryotes) is less clear. Lateral gene transfers (LGTs) have been argued to be potential sources of phylogenetic information, particularly for reconstructing deep nodes that are difficult to recover with traditional phylogenetic methods. We sought to identify interdomain LGTs by using a phylogenomic pipeline that generated 13 465 single gene trees and included up to 487 eukaryotes, 303 bacteria and 118 archaea. Our goals include searching for LGTs that unite major eukaryotic clades, and describing the relative contributions of LGT and EGT across the eukaryotic tree of life. Given the difficulties in interpreting single gene trees that aim to capture the approximately 1.8 billion years of eukaryotic evolution, we focus on presence–absence data to identify interdomain transfer events. Specifically, we identify 1138 genes found only in prokaryotes and representatives of three or fewer major clades of eukaryotes (e.g. Amoebozoa, Archaeplastida, Excavata, Opisthokonta, SAR and orphan lineages). The majority of these genes have phylogenetic patterns that are consistent with recent interdomain LGTs and, with the notable exception of EGTs involving photosynthetic eukaryotes, we detect few ancient interdomain LGTs. These analyses suggest that LGTs have probably occurred throughout the history of eukaryotes, but that ancient events are not maintained unless they are associated with endosymbiotic gene transfer among photosynthetic lineages.

1. Inferences about lateral and endosymbiotic gene transfer

The impact of lateral gene transfer (LGT) is best known in bacteria where the phenomenon of the rapid spread of antibiotic resistance among bacterial strains/species highlights the importance of this process in our own lives [1,2]. Analyses of first single genes and more recently whole genomes have demonstrated large numbers of LGTs among bacteria and archaea [3–6] and have contributed to discussions on the nature of species in these clades [7–9]. Less clear is the role of LGT in the evolution of eukaryotes, which may result from the use of the animals as models for evolutionary principles in eukaryotes given that the sequestration of the germline in triploblastic animals probably created a barrier to LGT [10–14]. There is a growing literature on LGTs involving eukaryotes, including microbial lineages [4,5,13], fungi [15,16], plants [17,18] and some animal lineages [19,20]. The bulk of these analyses focus on what might be termed 'tip down' approaches, asking about the impact of LGTs within clades rather than across the eukaryotic tree of life.

© 2015 The Author(s) Published by the Royal Society. All rights reserved.

In contrast to the debate on the role of LGT in eukaryotes, the past few decades have seen a substantial rise in descriptions of cases of endosymbiotic gene transfer (EGT): gene transfer from mitochondrion or plastid to the nucleus [21–24]. For example, roughly 15–20% of plant genomes are probably derived from plastid genes [25,26], and somewhere between 10% and 50% of the mitochondrial proteome is derived from nuclear encoded genes of alpha-proteobacterial origin [27,28]. Hence, transfer of genes from organelle to nucleus within a lineage is now well established [22,23,29,30].

Arguments have been made that LGTs can be used as evidence for ancient relationships, and that such data can be useful in reconstructing ancient relations [31,32]. For example, several authors have argued that there was a pulse of LGTs from various bacteria, including Chlamydiae, to the last common ancestor of Archaeplastida [33,34]. Analyses of networks generated by shared LGTs can be informative in discerning shared history among bacteria and/or archaea (hereafter termed prokaryotes) [35,36]. Abby *et al.* [35] use the distribution of LGT events in reconstructing relationships among bacteria and archaea, whereas Szollosi *et al.* [36] demonstrate the power of using LGTs to reconstruct the pattern and timing of events within bacterial genera and species.

Inferring ancient LGTs can be very difficult [22,37], which is why we focus on presence–absence data consistent with interdomain LGT. Inferring the transfer of single genes among divergent lineages of eukaryotes based on the topology of single gene trees is challenging given that we would be asking approximately 200–300 amino acids to estimate events as old as approximately 1.8 billion years (an estimate of the timing of eukaryotic origins [38,39]). In addition, errors in phylogenetic reconstruction such as long branch attraction and incomplete taxon sampling can mislead interpretations of lateral events based on tree topologies [22,40,41]. Perhaps most importantly for the analyses presented here, the prevalence of gene loss over evolutionary time [42] confounds interpretation of lateral events as genes present in bacteria plus only a few non-sister eukaryotic lineages may have been lost in other eukaryotic lineages. In other words, parallel gene loss among disparate lineages can mistakenly be interpreted as LGT among lineages retaining any given gene.

2. Estimating interdomain lateral gene transfers across the eukaryotic tree of life

Mindful of the perils and pitfalls of interpreting ancient gene transfer events, we set out to assess the tempo of interdomain LGTs and EGTs in the early evolution of eukaryotes by focusing on presence–absence data. Such analyses are possible because of our development of a phylogenomic pipeline that focuses on inclusion of a broad diversity of microbial lineages [43,44]. In brief, we start with clusters of homologous sequences (i.e. genes) as determined in OrthoMCL [45,46] and then add diverse taxa to end up with a sample of up to 487 eukaryotes, 303 bacteria and 118 archaea. The pipeline uses custom Python scripts and third-party tools such as Needle [47] to remove sequences that are either too similar (e.g. alleles) or too divergent (e.g. poor-quality transcripts, sequences sharing only motifs). Multi-sequence alignments are generated and refined using Guidance [48] and single gene trees are constructed using RaxML [49,50] using parameters from Grant & Katz [44,51]. We then used custom scripts to identify orthologues present in at least 10 taxa, at least three of which are bacteria/archaea, and a monophyletic clade of two or more eukaryotic sequences. Because of our focus on interdomain LGTs, the specifics of the resulting tree topologies (i.e. relationships among eukaryotes or among prokaryotes) are not critical, though we did require that eukaryotic sequences form a monophyletic clade.

Analyses of 13 465 genes, which included up to 908 diverse lineages (table 1 and electronic supplementary material, table S1), yielded 1138 genes that met our criteria for possible examples of interdomain transfers between prokaryotes and eukaryotes (table 2). We identified interdomain gene transfer events based on the presence of genes in prokaryotes plus members of three or fewer major eukaryotic clades (e.g. Opisthokonta, Archaeplastida; table 1) that formed a monophyletic group in the RaxML trees automatically generated by the pipeline [44]. This distribution was chosen because we think it is likely that orthologues present in four or more of the major eukaryotic clades were probably present in the last eukaryotic common ancestor (LECA).

We then categorized the 1138 genes based on their distributions in major (MC) and minor (mc) eukaryotic clades. Under this notation, genes in the 1MC1mc category are found in prokaryotes plus only one minor clade in only one major clade of eukaryotes, whereas genes in 3MC2mc are found in three major clades of eukaryotes with at least two minor clades in one of these major clades (table 2; electronic supplementary material, table S2). Because we controlled the names of our taxa to reflect major and minor clades (table 1; electronic supplementary material, table S1; [43,44]), categorizing genes was readily accomplished using the P4 package (https://code.google.com/p/p4-phylogenetics/). We also inspected the resulting trees to determine if a single or monophyletic minor clade of bacteria or archaea were sister to the eukaryotic sequences (electronic supplementary material, table S2), though we recognize the caution needed in interpreting these relationships given the likelihood of prokaryote–prokaryote LGT transfer [52].

3. The majority of putative interdomain lateral gene transfers appear to be recent events

Inspection of the 1138 genes that match our criteria for putative interdomain LGT reveal a striking pattern as over half of the putative interdomain LGTs (606 of 1138) involve only one minor clade nested within one major clade of eukaryotes (e.g. metazoa (Op_me) or ciliates (Sr_ci); table 2 and electronic supplementary material, S2). In fact, the greatest number of interdomain LGTs in this category are found in only one minor clade within the Opisthokonta (290 genes), Archaeplastida (170 genes) and then SAR (Stramenopila + Alveolata + Rhizaria; 59 genes; table 3). To exemplify this pattern, we include an example of one of the resulting trees for a putative carboxymuconolactone decarboxylase enzyme (OG5_141348 from OrthoMCL), which is found only in bacteria, archaea and animals (figure 1). The large number of interdomain LGTs into fungi (196 inferred; figure 2; electronic supplementary material, table S1) is consistent with numerous studies [15,53], but a broader comparison of our data with published cases of putative LGT is not easily done as we used more restrictive criteria (i.e. eukaryotes must be monophyletic) than most.

Table 1. Taxon sampling and major_minor clade abbreviations used in the analyses for the 487 eukaryotes, 303 bacteria and 118 archaea. Individual species/strain names are found in the electronic supplementary material, table S1. *n*, number of species/strains included in each category. Naming system is based largely on NCBI taxonomy, though no assumption is made on equivalency of rank for major (first abbreviation) and minor (second abbreviation) clades. The five major clades of eukaryotes each have a unique code (Op, Opisthokonta; Am, Amoebozoa; Ex, Excavata; Pl, Archaeplastida (Plantae); Sr, SAR (Stramenopila + Alveolata + Rhizaria)) and we use the abbreviation EE (everything else) to capture the non-monophyletic 'orphan' lineages (table 1).

eukaryotes		*n*	archaea/bacteria		*n*
Am_ac	Amoebozoa: Acanthamoebidae	1	Ar_cr	archaea: Crenarchaeota	27
Am_ar	Amoebozoa: Archamoebae	5	Ar_e	archaea: Euryarchaeota	77
Am_da	Amoebozoa: Discosea	1	Ar_ko	archaea: Korarchaeota	1
Am_di	Amoebozoa: Dictyostellida	3	Ar_na	archaea: Nanoarchaeota	1
Am_fi	Amoebozoa: Filamoeba	1	Ar_nh	archaea: Nanohaloarchaeota	3
Am_is	Amoebozoa: incertae sedis	4	Ar_pa	archaea: Parvarchaeota	2
Am_my	Amoebozoa: Mycetozoa	2	Ar_th	archaea: Thaumarchaeota	7
Am_va	Amoebozoa: Vannellidae	2	Ba_ac	bacteria: Actinobacteria	31
EE_ap	orphan: Apusozoa	1	Ba_ad	bacteria: Acidobacteria	1
EE_br	orphan: Breviatea	2	Ba_aq	bacteria: Aquificae	5
EE_cr	orphan: Cryptophyta	13	Ba_ar	bacteria: Armatimonadetes	1
EE_ha	orphan: Haptophyceae	16	Ba_ba	bacteria: Bacteroidetes	21
EE_is	orphan: incertae sedis	6	Ba_bc	bacteria: Chlorobi	3
EE_ka	orphan: Katablepharidophyta	1	Ba_bi	bacteria: Ignavibacteriae	2
Ex_eu	Excavata: Euglenozoa	23	Ba_ca	bacteria: Caldiserica	1
Ex_fo	Excavata: Fornicata	6	Ba_cd	bacteria: Chlamydiae	6
Ex_he	Excavata: Heterolobosea	4	Ba_ch	bacteria: Chloroflexi	9
Ex_is	Excavata: incertae sedis	1	Ba_cr	bacteria: Chrysiogenetes	1
Ex_ja	Excavata: Jakobida	5	Ba_cv	bacteria: Verrucomicrobia	4
Ex_ma	Excavata: Malawimonadidae	2	Ba_cy	bacteria: Cyanobacteria	41
Ex_ox	Excavata: Oxymonadida	1	Ba_de	bacteria: Deinococci	7
Ex_pa	Excavata: Parabasalia	4	Ba_df	bacteria: Deferribacteres	2
Op_ch	Opisthokonta: Choanoflagellida	5	Ba_di	bacteria: Dictyoglomi	2
Op_fu	Opisthokonta: fungi	40	Ba_el	bacteria: Elusimicrobia	1
Op_ic	Opisthokonta: Ichthyosporea	3	Ba_fb	bacteria: Firmicute, Bacilli	17
Op_is	Opisthokonta: incertae sedis	1	Ba_fc	bacteria: Firmicute, Clostridia	13
Op_me	Opisthokonta: Metazoa	61	Ba_fn	bacteria: Firmicute, Negativicutes	2
Pl_gl	Archaeplastida: Glaucocystophytes	3	Ba_fu	bacteria: Fusobacteria	5
Pl_gr	Archaeplastida: green algae	61	Ba_ge	bacteria: Gemmatimonadetes	1
Pl_rh	Archaeplastida: Rhodophyta	20	Ba_is	bacteria: incertae sedis	2
Sr_ap	SAR: Apicomplexa	18	Ba_me	bacteria: Melainabacteria	1
Sr_ch	SAR: Chromerida	2	Ba_ni	bacteria: Nitrospirae	3
Sr_ci	SAR: Ciliophora	27	Ba_pa	bacteria: Alphaproteobacteria	24
Sr_di	SAR: Dinophyceae	33	Ba_pb	bacteria: Betaproteobacteria	17
Sr_is	SAR: incertae sedis	1	Ba_pd	bacteria: Deltaproteobacteria	14
Sr_pe	SAR: Perkinsida	2	Ba_pg	Bacteria: Gammaproteobacteria	31
Sr_rh	SAR: Rhizaria	22	Ba_pl	bacteria: Planctomycetes	6
Sr_st	SAR: Stramenopila	84	Ba_sp	bacteria: Spirochaetes	9
			Ba_sy	bacteria: Synergistetes	4
			Ba_te	bacteria: Tenericutes	7
			Ba_th	bacteria: Thermotogae	7
			Ba_ts	bacteria: Thermodesulfobacterium	2

Table 2. Number of genes (orthologous groups) analysed for this study by category, organized based on those found in one, two or three major clades (MC) and noting the number of minor clades (mc). Using a starting set of 13 465 alignments/trees, we selected genes that met our criteria of having at least 10 sequences, at least three of which are bacteria/archaea, and a monophyletic clade of eukaryotes based on output of phylogenomic pipeline [43,44]. #MC refers to the number of major clades, including the non-monophyletic orphans (table 1). #mc refers to number of minor clades: 1mc = only one minor clade; 2mc ≥ two minor clades in at least one major clade. Individual genes are listed in electronic supplementary material, table S2.

abbreviation	description	*n*
1MC1mc	involving only one major clade and one minor clade	606
1MC2mc	involving only one major clade and at least two minor clades	77
2MC1mc	involving two major clades and only one minor clade in each	250
2MC2mc	involving two major clades and at least two minor clades in one major clade	140
3MC1mc	involving three major clades and only one minor clade in each	15
3MC2mc	involving three major clades and at least two minor clades in one major clade	50
total number		1138

Table 3. Recent interdomain LGTs involving prokaryotes and a single major clade of eukaryotes. Minor clade refers to nested taxa within the five major eukaryotic clades such as fungi with Opisthokonta or Apicomplexa within SAR. Numbers in parentheses are the number of species sampled (see electronic supplementary material, table S1). Individual genes are listed in electronic supplementary material, table S2.

	Amoebozoa (24)	orphans (37)	Excavata (49)	Opisthokonta (111)	Archaeplastida (85)	SAR (195)
one minor clade	42	2	43	290	170	59
≥two minor clades	3	0	3	31	29	11
total	45	2	46	321	199	70

Multiple factors probably impact these patterns including the number of lineages sampled within major clades (111, 85 and 195 for Opisthokonta, Archaeplastida and SAR, respectively; table 3) and the nature of the data used in our pipeline (e.g. genome sequence versus RNAseq [43,44]). Also, many lineages within Opisthokonta and Archaeplastida have relatively large genomes (e.g. http://data.kew.org/cvalues/, http://www.genomesize.com/) which may increase the probability of retention of ancient LGTs. Interdomain LGT events seem to be underrepresented among lineages within the Excavata, even though the pipeline includes whole genome data from several genomes in this clade (e.g. the genera *Leishmania*, *Trypanosoma*, *Giardia*, *Trichomonas*). As Excavata with whole-genome sequences are nearly all parasites and may have experienced considerable gene loss, analyses of free-living Excavata are likely to reveal additional examples of interdomain LGTs in this major eukaryotic clade.

We inspected the five LGTs that appear to define major clades (see asterisk symbol in figure 2) and, given the caveats discussed in the following, propose these be considered as only candidate synapomorphies until additional diverse lineages of eukaryotes are sampled. For example, the one gene that is found in at least three lineages of Opisthokonta (OG5_146700) is a hypothetic protein present in our pipeline in a subset of archaea, fungi, choanoflagellates and only one metazoan (electronic supplementary material, table S2). The three genes that may serve as synapomorphies for the SAR clade are patchily distributed (electronic supplementary material, table S2) and have diverse functions: a putative pyruvoyl tetrahydropterin synthase (OG5_141276), a penicillin amidase family protein (OG5_136942) and a putative deoxyribodipyrimidine photolyase (OG5_168036). Perhaps most optimistic as a synapomorphy is the one LGT at the base of Excavata, an acyl-CoA synthetase (OG5_146682), which has a relatively broad distribution given our sampling; it is found in Fornicata (*Giardia* and *Spironucleus*), Parabasalia (*Trichomonas*), Heterolobosea (*Sawyeria*) and Euglenozoa (*Euglena*).

4. The bulk of putative ancient interdomain gene transfers are likely to be endosymbiotic gene transfers

In contrast to the many recent interdomain LGTs, we see no compelling evidence for a pulse of LGT events that occurred in the common ancestors of major eukaryotic clades with the exception of gene transfers shared among clades with many photosynthetic members (table 4; electronic supplementary material, table S2; figure 2). The greatest numbers of putative interdomain LGTs involve clades with predominantly photosynthetic lineages (e.g. Archaeplastida (e.g. red and green algae), SAR (e.g. dinoflagellates, stramenopiles) and the 'orphan' Cryptophyta and Haptophyta (table 4 and electronic supplementary material, table S2). For example, there are 106 genes present in prokaryotes plus Archaeplastida plus SAR (table 4). Of the total of 455 genes that unite either two or three major clades of eukaryotes, 64 have gene tree topologies where photosynthetic eukaryotes are sister to cyanobacteria (electronic supplementary material, table S2). Retaining

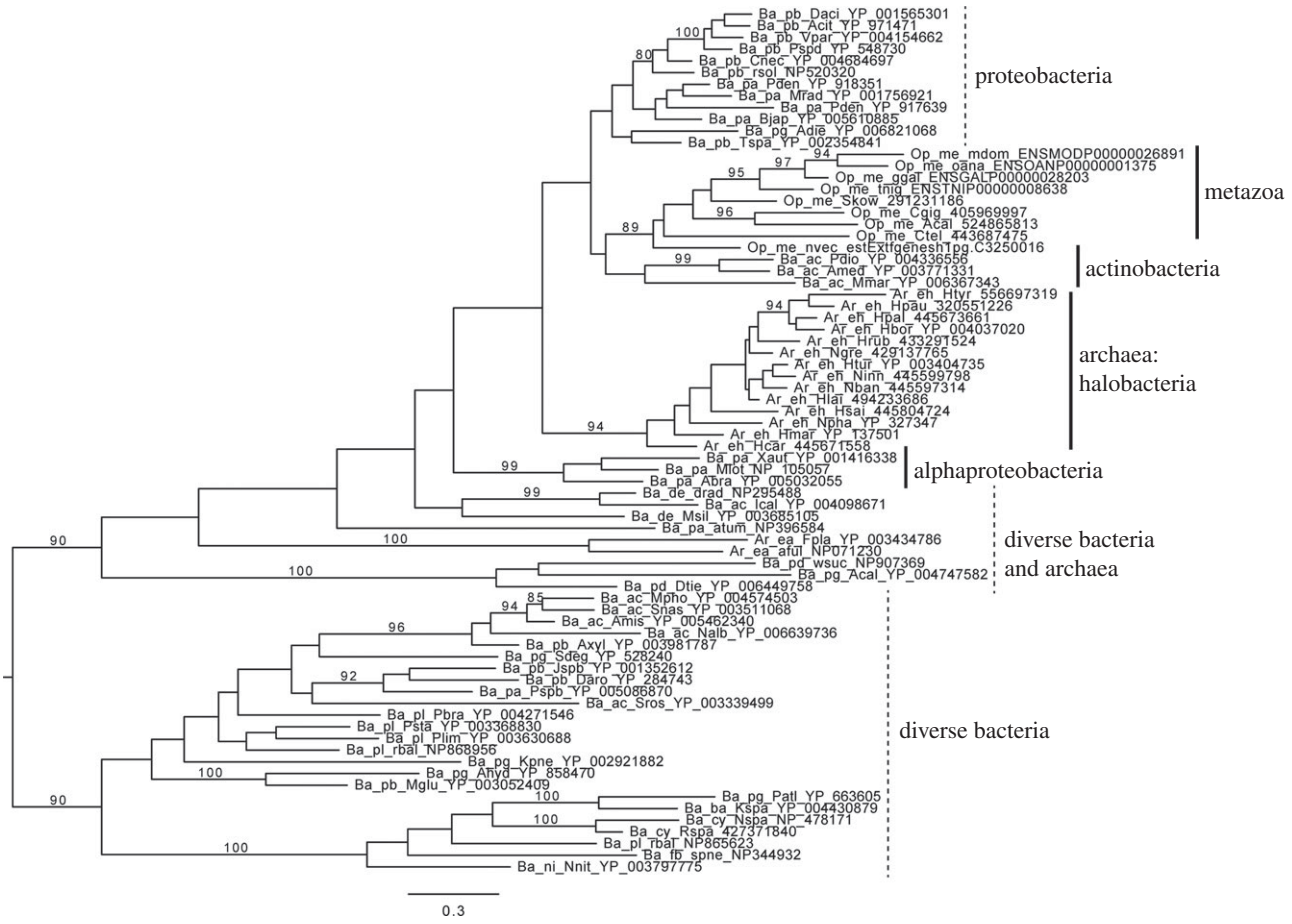


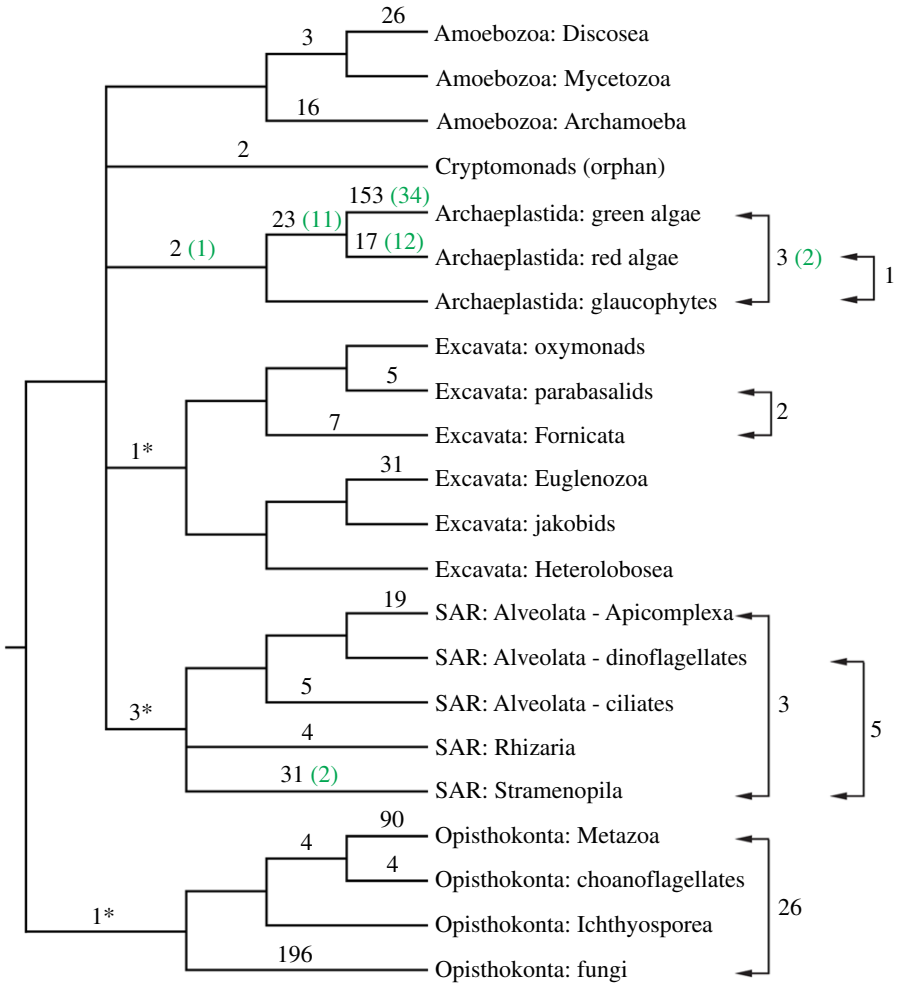
Figure 1. An example of a recent interdomain LGT from prokaryotes to one minor clade (Metazoa) in one major clade (Opisthokonta) of eukaryotes. This tree exemplifies the many recent (e.g. 1MC1mc) interdomain transfers detected in this study (table 2 and figure 3). Abbreviations of taxa are as in table 1 and electronic supplementary material, S1, and the number following each name is a unique identifier from either OrthoMCL or GenBank. Analyses of this gene used PROTGAMMA, the best-fitting LG model and default parameters as implemented in RAxML [49,50]. Most nodes are poorly supported and only bootstrap values above 80% are shown. Monophyletic clades are marked with solid lines, whereas the complex relationships among prokaryotes in the dashed clades probably represent a combination of poorly resolved phylogeny, LGT among prokaryotes and gene loss.

cyanobacterial sisters is neither a strict requirement nor predictor of EGT as subsequent LGTs among prokaryotes in the approximately 1 billion years since the acquisition of plastids may confound the signature of EGT [38]; nevertheless, these 64 gene trees provide additional support for EGT.

While there are numerous other genes that unite two (tables 3 and 4) or three (electronic supplementary material, table S2) major clades, there are no clear patterns except that the highest numbers of putative events are among clades with many species sampled and/or lineages with large genomes (table 3 and electronic supplementary material, table S2). In other words, interdomain LGTs do not appear to be a good source for synapomorphies for deep eukaryotic relationships. With the exception of events that unite photosynthetic lineages, we suspect that the putative LGT events counted in table 4 are either interdomain LGTs followed by loss or a combination of interdomain and intradomain LGTs. For example, a gene found only in bacteria plus two major clades of eukaryotes could be the result of (i) vertical ancestry plus loss in the remaining major clades or (ii) interdomain LGT followed by intradomain LGT. Discerning between such hypotheses is very challenging given the limited power within single gene trees.

To assess whether shared gene transfer events unite major clades of eukaryotes, we also used the software Coevolution Of Presence–Absence Patterns (CoPAP) [54]. COPAP is designed to detect patterns of co-evolving genes in presence–absence data from diverse lineages and uses efficient probabilistic models to assess the significance of relationships [54]. We inverted our data to ask whether there are taxa that share significant numbers of LGT events from putative cases of interdomain gene transfers. We used a p-value of 0.05 as cut-off for interactions, a star phylogeny for relationships among the 455 genes that were found in two or three major eukaryotic clades (2MC or 3MC; electronic supplementary material, table S2), plus default parameters as implemented by COPAP (http://copap.tau.ac.il/).

Only a small number of significant interactions are supported from analyses of presence–absence of 456 genes shared among two or three major clades of eukaryotes (figure 3). One significant network contains the predominantly photosynthetic lineages of dinoflagellates (Sr_di), glaucophytes (Pl_gl), red algae (Pl_rh), cryptophytes (EE_cr) and haptophytes (EE_ha). A second network contains the predominantly photosynthetic stramenopiles (Sr_st) plus green algae (Pl_gr), though this network is disconnected from the other photosynthetic lineages (figure 3). Uniting photosynthetic lineages is consistent with EGT from plastid to nucleus. Such transfers involve members of the Archaeplastida, the lineage descended from an ancestor that had a primary acquisition of plastids from cyanobacteria [58], plus the remaining lineages of photosynthetic eukaryotes (e.g. diatoms, brown algae, cryptophytes,



*= found in at least three minor clades

Figure 2. Recent LGT events mapped onto representative lineages from the eukaryotic tree of life. Numbers at nodes represent the LGT events in table 3, and the synthetic tree is arbitrarily rooted on Opisthokonta. Numbers marked by asterisk are found in at least three minor clades within major clades and may represent synapomorphies for major clades. Arrows on the right mark shared putative LGTs found between non-sister minor clades. Numbers in green (grey) in parentheses are genes where eukaryotes fall sister to cyanobacteria and are hence putative EGTs. For simplicity, only a subset of lineages are included here and full taxonomic distributions can be found in the electronic supplementary material, tables S1 and S2. (Online version in colour.)

Table 4. Putative interdomain transfers shared between two major eukaryotic clades. Parentheses contain the number of putative interdomain LGTs present in only one minor clade of both major clades, which suggests these may not be shared ancient events. Individual genes are listed in electronic supplementary material table S2.

	orphans	Excavata	Opisthokonta	Archaeplastida	SAR
Amoebozoa	1 (1)	11 (11)	28 (18)	9 (8)	8 (5)
orphans	—	0	7 (1)	17 (13)	12 (6)
Excavata		—	26 (21)	8 (7)	9 (4)
Opisthokonta			—	71 (58)	76 (41)
Archaeplastida				—	106 (55)
SAR					—

haptophytes and dinoflagellates) that acquired photosynthesis through secondary endosymbiosis [59,60].

The only other significant cluster detected by COPAP is the pairing of the archamoeba parasites *Entamoeba* spp. (Am_ar) with parabasalids including *Trichomonas* (Ex:pa; figure 3). An association between *Entamoeba* spp. and parabasalids was first described by several authors [55–57] before we found support of this hypotheses through analyses of interdomain and intradomain LGT involving *Entamoeba* spp. [51]. Given that we are using the same dataset here,

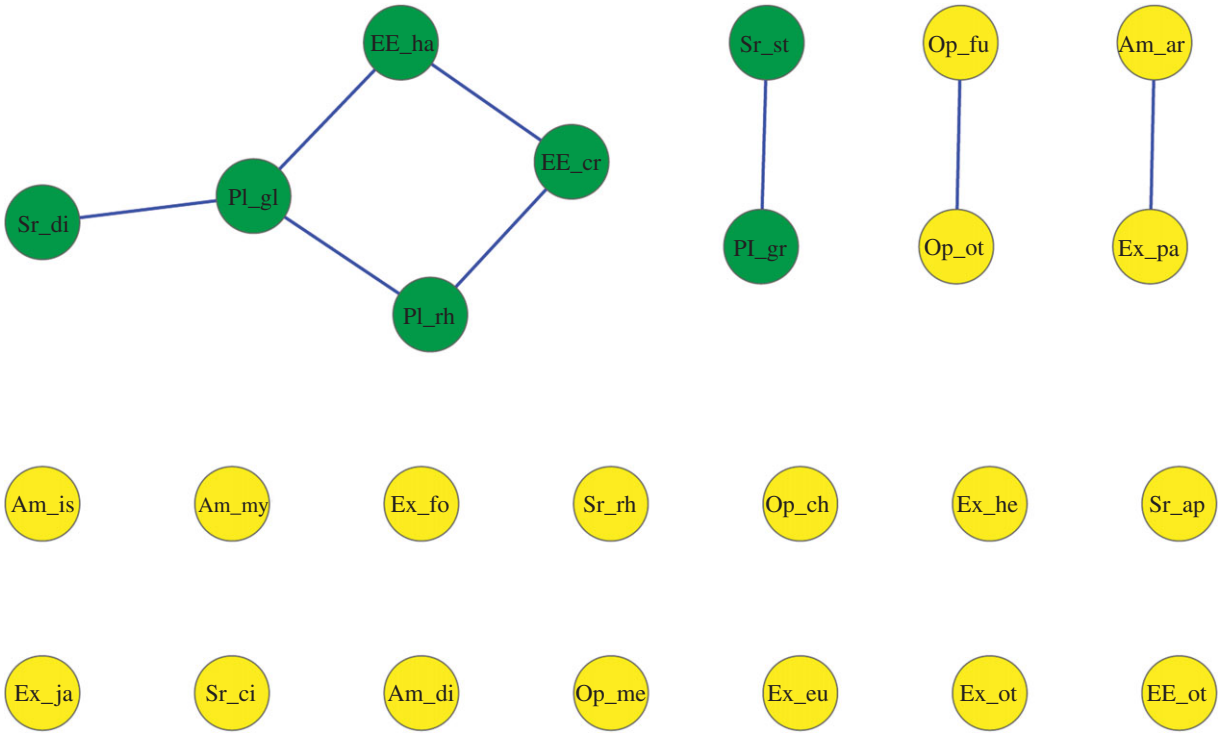


Figure 3. Significant networks among lineages as determined by COPAP [54] based on presence–absence of LGTs. The green (dark) taxa are predominantly photosynthetic, and networks involving these minor clades indicate the potential influence of EGT on photosynthetic lineages. There is also a significant relationship between *Entamoeba* spp. (Am_ar) and parabasalids (Ex_pa) as has been previously observed [51,55–57]. The linking of fungi (Op_fu) and microbial opisthokonts (Op_ot; other Opisthokonta = Ichthyosporea plus lineages that are incertae sedis) probably represents shared retention of LGT events. (Online vesion in colour.)

our finding this association is not surprising, though COPAP evaluates patterns of gene presence–absence as opposed to tree topologies.

5. Numerous caveats must be considered when interpreting patterns of lateral gene transfers

There are many caveats to be considered when interpreting ancient gene transfer events. Insights on both EGT and LGT are dependent on taxon sampling, which is uneven in our dataset in terms of the availability and quality of data from diverse lineages. The impact of taxon sampling on inferences about ancient LGT can be seen in the changing narrative about a single gene transfer of a tyrosyl-tRNA synthetase gene, which was originally argued to be a synapomorphy for Opisthokonta based on available data [61]. Reanalysis with data from additional microbial eukaryotes revealed that this gene was also present in Amoebozoa [40]. In our expanded taxon sampling, we find this gene in multiple lineages in Amoebozoa plus the parasite *Blastocystis homoni* (Sr_st), the orphan lineage *Palpitomonas bilix* (EE_is_Pbil) and two Rhizarian species (Sr_rh; electronic supplementary material, figure S1). While additional work is needed to rule out that these additional sparsely distributed taxa are not spurious data (i.e. contamination), our changing understanding of the phylogenetic distribution of the tyrosyl-tRNA synthetase gene highlights the impact of taxon sampling on inferences of ancient gene transfer events.

We also checked to see whether we find a pattern of interdomain LGT from Chlamydiae to Archaeplastida and other photosynthetic eukaryotes, which was observed in several previous analyses [33,34,62,63]. Only 10 genes of our 1138 matched the criteria of being present in three or fewer major eukaryotic clades with a single species or monophyletic clade of Chlamydiae as sister taxa (electronic supplementary material, table S2). Of these, only one (OG5_146631, an FAD-dependent oxidoreductase family protein) is found exclusively in photosynthetic eukaryotes. We also looked at the gene trees created by our pipeline for the 38 proteins reported as possible cases of transfer from Chlaymdiae to photosynthetic eukaryotes, and found that only half could be argued to be consistent with this transfer hypothesis given our taxonomic sampling (electronic supplementary material, table S3). Importantly, only two of the 38 genes reported by Becker *et al*. [34] match the more conservative criteria employed in our analyses as eukaryotes are not monophyletic in the remaining 36 trees. For our analyses of interdomain LGTs, we rejected trees with non-monophyletic eukaryotes as we do not believe we can distinguish between multiple interdomain LGTs and poorly resolved phylogenies without more in depth phylogenetic analyses.

Gene loss is clearly a major force in genome evolution [13,42], so interpreting the ancient LGT events must be done with caution particularly given the variation in genome sizes among the eukaryotes sampled for this study. Based on inferences on patterns of intron loss in eukaryotes and on genome complexity in the LECA, Wolf & Koonin [42] argue that gene loss has dominated the evolution of eukaryotic genomes, with intervening periods of 'complexification' that may include pulses of LGT/EGT. Because of the importance of gene loss we recognize that some of our examples of recent interdomain LGT may instead be more ancient gene transfer events that were then lost in major clades of eukaryotes. At the same time, we anticipate that even more recent events will be found as taxon sampling

expands, particularly in poorly sampled territories like much of the Excavata and Rhizaria [43].

6. Conclusion

We identify 1138 genes that meet our criteria of possible interdomain LGTs as they are found in prokaryotes plus three or fewer major clades of eukaryotes. Analyses of the patterns among these genes reveals evidence of recent interdomain LGT events between prokaryotes and eukaryotes (table 3 and figure 2) and no compelling evidence of retained ancient LGTs (i.e. those that occurred in eukaryotic ancestors prior to the divergence of major clades). In contrast, we do detect numerous examples of EGTs involving multiple lineages of photosynthetic eukaryotes (figure 2 and figure 3), which validates our phylogenomic approach to detecting interdomain gene transfer events as the impact of EGTs has been established using other approaches [21–23,29]. With the exception of the EGTs, the data presented here are consistent with a model whereby gene loss is a dominant force in the evolution of eukaryotic genomes [42] as our analyses indicate that most interdomain LGTs have been lost over evolutionary time.

Note added in proof

The findings reported here are generally concordant with those from analyses of approximately 100 000 genes in a more limited number of eukaryotes plus prokaryotes (Ku *et al*. In press. Nature).

Data accessibility. Alignments and trees for the 1138 genes are available in the Dryad Digital Repository (http://dx.doi.org/10.5061/dryad.2bj36).

Competing interests. We declare we have no competing interests.

Funding. The work here was supported by NSF grant DEB-1208741 and NIH award 1R15GM113177.

Acknowledgements. The analyses were done with the help of Jessica R. Grant, a research associate at Smith College who developed and implemented the phylogenomic pipeline to collect the data for this manuscript. The manuscript benefited from comments of two anonymous reviewers as well as from the undergraduate and graduate students working in the Katz laboratory.

References

1. Juhas M. 2015 Horizontal gene transfer in human pathogens. *Crit. Rev. Microbiol*. **41**, 101–108. (doi:10.3109/1040841x.2013.804031)
2. Palmer KL, Kos VN, Gilmore MS. 2010 Horizontal gene transfer and the genomics of enterococcal antibiotic resistance. *Curr. Opin. Microbiol*. **13**, 632–639. (doi:10.1016/j.mib.2010.08.004)
3. Le PT, Pontarotti P, Raoult D. 2014 Alphaproteobacteria species as a source and target of lateral sequence transfers. *Trends Microbiol*. **22**, 147–156. (doi:10.1016/j.tim.2013.12.006)
4. Dagan T. 2011 Phylogenomic networks. *Trends Microbiol*. **19**, 483–491. (doi:10.1016/j.tim.2011.07.001)
5. Kloesges T, Popa O, Martin W, Dagan T. 2011 Networks of gene sharing among 329 proteobacterial genomes reveal differences in lateral gene transfer frequency at different phylogenetic depths. *Mol. Biol. Evol*. **28**, 1057–1074. (doi:10.1093/molbev/msq297)
6. Doolittle WF, Bapteste E. 2007 Pattern pluralism and the tree of life hypothesis. *Proc. Natl Acad. Sci. USA* **104**, 2043–2049. (doi:10.1073/pnas.0610699104)
7. Barraclough TG, Balbi KJ, Ellis RJ. 2012 Evolving concepts of bacterial species. *Evol. Biol*. **39**, 148–157. (doi:10.1007/s11692-012-9181-8)
8. Konstantinidis KT, Ramette A, Tiedje JM. 2006 The bacterial species definition in the genomic era. *Phil. Trans. R. Soc. B* **361**, 1929–1940. (doi:10.1098/rstb.2006.1920)
9. Caro-Quintero A, Konstantinidis KT. 2012 Bacterial species may exist, metagenomics reveal. *Environ. Microbiol*. **14**, 347–355. (doi:10.1111/j.1462-2920.2011.02668.x)
10. Boto L. 2010 Horizontal gene transfer in evolution: facts and challenges. *Proc. R. Soc. B* **277**, 819–827. (doi:10.1098/rspb.2009.1679)
11. Huang J. 2013 Horizontal gene transfer in eukaryotes: the weak-link model. *BioEssays* **35**, 868–875. (doi:10.1002/bies.201300007)
12. Katz LA. 2002 Lateral gene transfers and the evolution of eukaryotes: theories and data. *Int. J. Syst. Evol. Microbiol*. **52**, 1893–1900. (doi:10.1099/ijs.0.02113-0)
13. Keeling PJ, Palmer JD. 2008 Horizontal gene transfer in eukaryotic evolution. *Nat. Rev. Genet*. **9**, 605–618. (doi:10.1038/Nrg2386)
14. Schönknecht G, Weber APM, Lercher MJ. 2014 Horizontal gene acquisitions by eukaryotes as drivers of adaptive evolution. *BioEssays* **36**, 9–20. (doi:10.1002/bies.201300095)
15. Zhang MZ, Silva M, Maryam CD, van Elsas JD. 2014 The mycosphere constitutes an arena for horizontal gene transfer with strong evolutionary implications for bacterial–fungal interactions. *FEMS Microbiol. Ecol*. **89**, 516–526. (doi:10.1111/1574-6941.12350)
16. Fitzpatrick DA. 2012 Horizontal gene transfer in fungi. *FEMS Microbiol. Lett*. **329**, 1–8. (doi:10.1111/j.1574-6968.2011.02465.x)
17. Gao CH, Ren XD, Mason AS, Liu HL, Xiao ML, Li JN, Fu DH. 2014 Horizontal gene transfer in plants. *Funct. Integr. Genomics* **14**, 23–29. (doi:10.1007/s10142-013-0345-0)
18. Yue JP, Hu XY, Sun H, Yang YP, Huang JL. 2012 Widespread impact of horizontal gene transfer on plant colonization of land. *Nat. Commun*. **3**, 1152. (doi:10.1038/ncomms2148)
19. Boto L. 2014 Horizontal gene transfer in the acquisition of novel traits by metazoans. *Proc. R. Soc. B* **281**, 20132450. (doi:10.1098/rspb.2013.2450)
20. Degnan SM. 2014 Think laterally: horizontal gene transfer from symbiotic microbes may extend the phenotype of marine sessile hosts. *Front. Microbiol*. **5**, 638. (doi:10.3389/fmicb.2014.00638)
21. Meyer-Gauen G, Schnarrenberger C, Cerff R, Martin W. 1994 Molecular characterization of a novel, nuclear-encoded, NAD^+-dependent glyceraldehyde-3-phosphate dehydrogenase in plastids of the gymnosperm *Pinus sylvestris* L. *Plant Mol. Biol*. **26**, 1155–1166. (doi:10.1007/BF00040696)
22. Ku C, Nelson-Sathi S, Roettger M, Garg S, Hazkani-Covo E, Martin WF. In press. Endosymbiotic gene transfer from prokaryotic pangenomes: inherited chimerism in eukaryotes. *Proc. Natl Acad. Sci. USA* (doi:10.1073/pnas.1421385112)
23. Timmis JN, Ayliffe MA, Huang CY, Martin W. 2004 Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes. *Nat. Rev. Genet*. **5**, 123–135. (doi:10.1038/nrg1271)
24. Gabaldon T, Huynen MA. 2005 Lineage-specific gene loss following mitochondrial endosymbiosis and its potential for function prediction in eukaryotes. *Bioinformatics* **21**(Suppl. 2), 144–150. (doi:10.1093/bioinformatics/bti1124)
25. Martin W, Rujan T, Richly E, Hansen A, Cornelsen S, Lins T, Leister D, Stoebe B, Hasegawa M, Penny D. 2002 Evolutionary analysis of *Arabidopsis*, cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus. *Proc. Natl Acad. Sci. USA* **99**, 12 246–12 251. (doi:10.1073/pnas.182432999)
26. Deusch O, Landan G, Roettger M, Gruenheit N, Kowallik KV, Allen JF, Martin W, Dagan T. 2008 Genes of cyanobacterial origin in plant nuclear genomes point to a heterocyst-forming plastid ancestor. *Mol. Biol. Evol*. **25**, 748–761. (doi:10.1093/molbev/msn022)
27. Gray MW. 2014 The pre-endosymbiont hypothesis: a new perspective on the origin and evolution of mitochondria. *Cold Spring Harbor Perspect. Biol*. **6**, a016097. (doi:10.1101/cshperspect.a016097)

28. Gabaldon T, Huynen MA. 2004 Shaping the mitochondrial proteome. *Biochim. Biophys. Acta Bioenerg.* **1659**, 212–220. (doi:10.1016/j.bbabio.2004.07.011)
29. Qiu H, Price DC, Weber APM, Facchinelli F, Yoon HS, Bhattacharya D. 2013 Assessing the bacterial contribution to the plastid proteome. *Trends Plant Sci*. **18**, 680–687. (doi:10.1016/j.tplants.2013.09.007)
30. Gabaldon T, Huynen MA. 2007 From endosymbiont to host-controlled organelle: the hijacking of mitochondrial protein synthesis and metabolism. *PLoS Comput. Biol.* **3**, e219. (doi:10.1371/journal.pcbi.0030219)
31. Fournier GP, Huang J, Gogarten JP. 2009 Horizontal gene transfer from extinct and extant lineages: biological innovation and the coral of life. *Phil. Trans. R. Soc. Lond. B* **364**, 2229–2239. (doi:10.1098/rstb.2009.0033)
32. Williams D, Gogarten JP, Lapierre P. 2010 Filling the gaps in the genomic landscape. *Genome Biol*. **11**, 103. (doi:10.1186/gb-2010-11-2-103)
33. Huang JL, Gogarten JP. 2008 Concerted gene recruitment in early plant evolution. *Genome Biol*. **9**, R109. (doi:10.1186/gb-2008-9-7-r109)
34. Becker B, Hoef-Emden K, Melkonian M. 2008 Chlamydial genes shed light on the evolution of photoautotrophic eukaryotes. *BMC Evol. Biol*. **8**, 203. (doi:10.1186/1471-2148-8-203)
35. Abby SS, Tannier E, Gouy M, Daubin V. 2012 Lateral gene transfer as a support for the tree of life. *Proc. Natl Acad. Sci. USA* **109**, 4962–4967. (doi:10.1073/pnas.1116871109)
36. Szollosi GJ, Boussau B, Abby SS, Tannier E, Daubin V. 2012 Phylogenetic modeling of lateral gene transfer reconstructs the pattern and relative timing of speciations. *Proc. Natl Acad. Sci. USA* **109**, 17 513–17 518. (doi:10.1073/pnas.1202997109)
37. Andersson JO. 2009 Gene transfer and diversification of microbial eukaryotes. *Annu. Rev. Microbiol*. **63**, 177–193. (doi:10.1146/annurev.micro.091208.073203)
38. Parfrey LW, Lahr DJG, Knoll AH, Katz LA. 2011 Estimating the timing of early eukaryotic diversification with multigene molecular clocks. *Proc. Natl Acad. Sci. USA* **108**, 13 624–13 629. (doi:10.1073/Pnas.1110633108)
39. Knoll AH, Javaux EJ, Hewitt D, Cohen P. 2006 Eukaryotic organisms in Proterozoic oceans. *Phil. Trans. R. Soc. B* **361**, 1023–1038. (doi:10.1098/rstb.2006.1843)
40. Shadwick JDL, Ruiz-Trillo I. 2012 A genomic survey shows that the haloarchaeal type tyrosyl tRNA synthetase is not a synapomorphy of opisthokonts. *Eur. J. Protistol.* **48**, 89–93. (doi:10.1016/j.ejop.2011.10.003)
41. Almeida FC, Leszczyniecka M, Fisher PB, DeSalle R. 2008 Examining ancient inter-domain horizontal gene transfer. *Evol. Bioinf.* **4**, 109–119.
42. Wolf YI, Koonin EV. 2013 Genome reduction as the dominant mode of evolution. *BioEssays* **35**, 829–837. (doi:10.1002/bies.201300037)
43. Katz LA, Grant JR. 2015 Taxon-rich phylogenomic analyses resolve the eukaryotic tree of life and reveal the power of subsampling by sites. *Syst. Biol*. **64**, 406–415. (doi:10.1093/sysbio/syu126)
44. Grant JR, Katz LA. 2014 Building a phylogenomic pipeline for the eukaryotic tree of life: addressing deep phylogenies with genome-scale data. *PLoS Curr.* 2 April. Edition 1. (doi:10.1371/currents.tol.c24b6054aebf3602748ac042ccc8f2e9)
45. Chen F, Mackey AJ, Stoeckert CJ, Roos DS. 2006 OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. *Nucleic Acids Res*. **34**, D363–D368. (doi:10.1093/nar/gkj123)
46. Li L, Stoeckert CJJr, Roos DS. 2003 OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res* **13**, 2178–2189. (doi:10.1101/gr.1224503)
47. Rice P, Longden I, Bleasby A. 2000 EMBOSS: the European molecular biology open software suite. *Trends Genet*. **16**, 276–277. (doi:10.1016/S0168-9525(00)02024-2)
48. Penn O, Privman E, Ashkenazy H, Landan G, Graur D, Pupko T. 2010 GUIDANCE: a web server for assessing alignment confidence scores. *Nucleic Acids Res*. **38**, W23–W28. (doi:10.1093/nar/gkq443)
49. Stamatakis A, Hoover P, Rougemont J. 2008 A rapid bootstrap algorithm for the RAxML web-servers. *Syst. Biol*. **57**, 758–771. (doi:10.1080/10635150802429642)
50. Stamatakis A. 2006 RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22**, 2688–2690. (doi:10.1093/bioinformatics/btl446)
51. Grant JR, Katz LA. 2014 Phylogenomic study indicates widespread lateral gene transfer in *Entamoeba* and suggests a past intimate relationship with parabasalids. *Genome Biol. Evol.* **6**, 2350–2360. (doi:10.1093/gbe/evu179)
52. Dagan T, Martin W. 2009 Getting a better picture of microbial evolution en route to a network of genomes. *Phil. Trans. R. Soc. B* **364**, 2187–2196. (doi:10.1098/rstb.2009.0040)
53. Jaramillo VD, Sukno SA, Thon MR. 2015 Identification of horizontally transferred genes in the genus *Colletotrichum* reveals a steady tempo of bacterial to fungal gene transfer. *BMC Genomics* **16**, 2. (doi:10.1186/1471-2164-16-2)
54. Cohen O, Ashkenazy H, Karin EL, Burstein D, Pupko T. 2013 CoPAP: coevolution of presence–absence patterns. *Nucleic Acids Res*. **41**, W232–W237. (doi:10.1093/nar/gkt471)
55. Stanley SLJr. 2005 The *Entamoeba histolytica* genome: something old, something new, something borrowed and sex too? *Trends Parasitol.* **21**, 451–453. (doi:10.1016/j.pt.2005.08.006)
56. Alsmark UC, Sicheritz-Ponten T, Foster PG, Hirt RP, Embley TM. 2009 Horizontal gene transfer in eukaryotic parasites: a case study of *Entamoeba histolytica* and *Trichomonas vaginalis*. *Methods Mol. Biol*. **532**, 489–500. (doi:10.1007/978-1-60327-853-9_28)
57. Alsmark C, Foster PG, Sicheritz-Ponten T, Nakjang S, Embley TM, Hirt RP. 2013 Patterns of prokaryotic lateral gene transfers affecting parasitic microbial eukaryotes. *Genome Biol*. **14**, R19. (doi:10.1186/Gb-2013-14-2-R19)
58. Adl SM *et al.* 2005 The new higher level classification of eukaryotes with emphasis on the taxonomy of protists. *J. Eukaryot. Microbiol*. **52**, 399–451. (doi:10.1111/j.1550-7408.2005.00053.x)
59. Delwiche CF. 1999 Tracing the tread of plastid diversity through the tapestry of life. *Am. Nat*. **154**, S164–S177. (doi:10.1086/303291)
60. Archibald JM. 2009 The puzzle of plastid evolution. *Curr. Biol*. **19**, R81–R88. (doi:10.1016/j.cub.2008.11.067)
61. Huang JL, Xu Y, Gogarten JP. 2005 The presence of a haloarchaeal type tyrosyl-tRNA synthetase marks the opisthokonts as monophyletic. *Mol. Biol. Evol*. **22**, 2142–2146. (doi:10.1093/molbev/msi221)
62. Huang JL, Yue JP. 2013 Horizontal gene transfer in the evolution of photosynthetic eukaryotes. *J. Syst. Evol.* **51**, 13–29. (doi:10.1111/j.1759-6831.2012.00237.x)
63. Moustafa A, Reyes-Prieto A, Bhattacharya D. 2008 Chlamydiae has contributed at least 55 genes to plantae with predominantly plastid functions. *PLoS ONE* **3**, e2205. (doi:10.1371/journal.pone.0002205)