Failure to Recover Major Events of Gene Flux in Real Biological Data Due to Method Misapplication

Nils Kapust[1], Shijulal Nelson-Sathi[2], Barbara Schönfeld[3], Einat Hazkani-Covo[4], David Bryant[5], Peter J. Lockhart[6], Mayo Röttger[1], Joana C. Xavier[1,*], and William F. Martin[1]

[1]Institute of Molecular Evolution, Heinrich Heine University, Düsseldorf, Germany

[2]Computational Biology & Bioinformatics Group, Rajiv Gandhi Centre for Biotechnology, Trivandrum, Kerala, India

[3]School of Zoology, University of Tasmania, Hobart, Australia

[4]Department of Natural and Life Sciences, The Open University of Israel, Ra'anana, Israel

[5]Department of Mathematics and Statistics, University of Otago, Dunedin, New Zealand

[6]Institute of Fundamental Sciences, Massey University, Palmerston North, New Zealand

*Corresponding author: E-mail: xavier@hhu.de.

Accepted: April 17, 2018

Abstract

In prokaryotes, known mechanisms of lateral gene transfer (transformation, transduction, conjugation, and gene transfer agents) generate new combinations of genes among chromosomes during evolution. In eukaryotes, whose host lineage is descended from archaea, lateral gene transfer from organelles to the nucleus occurs at endosymbiotic events. Recent genome analyses studying gene distributions have uncovered evidence for sporadic, discontinuous events of gene transfer from bacteria to archaea during evolution. Other studies have used traditional models designed to investigate gene family size evolution (Count) to support claims that gene transfer to archaea was continuous during evolution, rather than involving occasional periodic mass gene influx events. Here, we show that the methodology used in analyses favoring continuous gene transfers to archaea was misapplied in other studies and does not recover known events of single simultaneous origin for many genes followed by differential loss in real data: plastid genomes. Using the same software and the same settings, we reanalyzed presence/absence pattern data for proteins encoded in plastid genomes and for eukaryotic protein families acquired from plastids. Contrary to expectations under a plastid origin model, we found that the methodology employed inferred that gene acquisitions occurred uniformly across the plant tree. Sometimes as many as nine different acquisitions by plastid DNA were inferred for the same protein family. That is, the methodology that recovered gradual and continuous lateral gene transfer among lineages for archaea obtains the same result for plastids, even though it is known that massive gains followed by gradual differential loss is the true evolutionary process that generated plastid gene distribution data. Our findings caution against the use of models designed to study gene family size evolution for investigating gene transfer processes, especially when transfers involving more than one gene per event are possible.

Key words: LGT, archaea, evolutionary models, plastid genomes.

Introduction

Lateral gene transfer (LGT) has had a major impact on gene distributions among archaeal chromosomes during evolution (Wagner et al. 2017). There are basically two ways that are currently employed to infer the evolutionary processes underlying gene distributions. One approach is to construct phylogenetic trees for all proteins in a given set of genomes and to compare topologies in search of phylogenetic congruence or incongruence, evoking vertical inheritance to account for the former and LGT to account for the latter. Despite the occurrence of historical events of lateral gene transfer among prokaryotes, applications of this approach have nevertheless generally led to phylogenetic reconstructions favoring a single dominant underlying prokaryotic tree (e.g., in Daubin et al. 2003). However, some tree building studies with prokaryotes have uncovered significant amounts of phylogenetic incongruence, also among early diverging lineages, and suggested lateral transfer of genes from bacteria to archaea (e.g., in Pisani et al. 2007).

Species tree reconstructions from multiple independent gene loci have been used to study gene distributions. This approach requires reconciliation of all gene tree topologies, all gene duplications, all gene losses, and all gene transfers simultaneously for a given set of species (Szöllősi, Davín et al. 2015). The trouble with this approach is that the number of model parameters is large and the processes of molecular evolution complex. Models with few parameters run the very real risk of systematic error from model misspecification. Models with large numbers of parameters make it difficult to draw precise conclusions and quantify statistical error, as was recently observed for analyses of gene phylogenies addressing mitochondrial origin (Martin, Roettger et al. 2017).

One limitation of tree-based approaches is that they are impacted by both incomplete taxon sampling and by the fact that the vast majority of genes in prokaryotes occur in only a very few genomes (Dagan and Martin 2007). Genes present in only two or three genomes will appear to have been vertically inherited in all trees, and at least a third of all genes present in four genomes will also appear to be vertically inherited by phylogenetic congruence criteria alone. The problem with this potential methodological bias is that it will inflate ancestral genome sizes to unacceptably large values if one looks at all genes (Dagan and Martin 2007), not just the ones for which trees are convenient to construct.

An alternative and still relatively new approach for studying gene distribution is to investigate the factors underlying gene distributions, to cluster all protein coding genes in a given set of genomes into protein families, and to examine not only the presence/absence patterns of those genes along a given reference tree but also the phylogenies for each individual cluster (Nelson-Sathi et al. 2012; Ku et al. 2015). When applied to archaea, this approach uncovered that haloarchaea acquired ∼1000 genes from bacteria in a process that transformed a chemolithoautotrophic methanogen ancestor into a facultative aerobic heterotroph (Nelson-Sathi et al. 2012) and that gene acquisitions from bacteria followed by extensive differential loss was important in the origin and evolution of several major archaeal clades (Nelson-Sathi et al. 2015). The same fundamental pattern is observed in eukaryote evolution, where the host lineage is thought to descend from archaea (Martin and Müller 1998; Williams et al. 2013; McInerney et al. 2014; Zaremba-Niedzwiedzka et al. 2017), namely events of mass gene acquisition followed by differential loss (Ku et al. 2015), loss being a very important factor in genome evolution (Albalat and Cañestro 2016).

Recently, Groussin et al. (2016) reanalyzed the data of Nelson-Sathi et al. (2015) using a program called Count (Csűrös 2010). Count implements a wide range of methods and models for reconstructing the evolution of gene family sizes within multiple species in a phylogeny. These include birth–death process models for gene gain and loss within families, and LGT between lineages. Groussin et al. (2016) made the curious decision to use Count (Csűrös 2010) to investigate binary (gene presence/absence) data, essentially shoe-horning binary data into analysis with a method that is designed to investigate numerical data (integers >1, changes in gene family size) across a phylogeny. Groussin et al. (2016) inferred essentially the same amount of LGT as Nelson-Sathi et al. (2015) found, but Count distributed the LGTs across the reference tree in such a way as to evenly distribute gains and losses according to the models of gene family counts implemented in the software. From that result, they concluded that LGT was mostly uniform and continuous during archaeal evolution (Groussin et al. 2016), not episodic (Nelson-Sathi et al. 2015). That however involved misapplication of the models in Count.

The same misapplication of the models in Count also leads to inference of vast amounts of continuous LGT during eukaryote evolution (Szöllősi, Tannier et al. 2015), even though there are no known genetic mechanisms for LGT among eukaryotes (Martin 2017), in contrast to the very well characterized mechanisms of LGT among prokaryotes (Popa and Dagan 2011). There are reasons to suspect that the amounts of LGT that Szöllősi, Tannier et al. (2015) found for fungi (eukaryotes) are methodological artefacts, because if eukaryotes were exchanging genes freely across higher taxonomical boundaries then eukaryote genomes should exhibit cumulative effects of LGT as prokaryote genomes do, but the converse is observed (Martin 2017). Moreover, genome-scale tests for eukaryote LGT show that gene evolution in eukaryotes is vertical, mediated by loss, and punctuated by gene acquisitions at endosymbiotic events (Ku et al. 2015; Ku and Martin 2016).

We suspected that model misspecification and model misapplication explain the unusual results obtained in the study by (Groussin et al. 2016), where Count was forced to analyze presence/absence data. However even if that problem was fixed, deeper challenges remain. A critical assumption in this context is that the evolutionary histories of different gene families are independent of one another. Thus, an LGT involving a transfer of *n* genes would be considered as *n* individual events, because each gene is considered individually in gene presence/absence studies. Major acquisition events, which alter the evolutionary fate of large numbers of genes in a relatively short period of time fall completely outside the scope of current models that investigate gene presence/absence data across a reference tree (Groussin et al. 2016). The method of Nelson-Sathi et al. (2015) does not investigate gene presence/absence patterns across a reference tree, it estimates the probability that two independent sets of trees with nonidentical leaf sets are drawn from the same distribution. Of further concern is the potential impact of imposing constant rates of gene birth and death across different gene families (Szöllősi, Tannier et al. 2015; Groussin et al. 2016). To examine the extent to which these systematic problems impact the inference of genome evolution, we apply the methodology used by Groussin et al. (2016) and Szöllősi, Tannier et al. (2015) to infer the distribution of plastid genes that were acquired via endosymbiosis and then lost during plastid genome evolution. For plastid genome data, we know in advance that the "loss only model" correctly describes the evolution of gene content.

Materials and Methods

Data Collection and Annotation

Archaeal Protein Families

The data set used for the study of the origin of archaeal protein families included 1,981 prokaryotic genomes—134 archaea and 1,847 bacteria (Nelson-Sathi et al. 2015), hereafter referred to as AR data set. The amino acid sequences were retrieved from RefSeq, NCBI (version June 2012). The data set consists of 254,938 archaeal proteins in 25,762 protein families, of which the subset consisting of the import clusters (13,631 archaeal proteins in 2,264 protein families), used in Groussin et al. (2016), was used as well here.

Plastid Protein Families

A data set encompassing all plastid encoded proteins for 193 photosynthetic eukaryotes (Schönfeld 2012), designated as the PL data set, was used. It consists of 254 protein families from 193 sequenced plastid genomes of different eukaryotes, encompassing 6,561 protein sequences in total. All sequences were retrieved from RefSeq, NCBI (version January 2011). Each protein family was manually annotated into Uniprot functional categories.

Eukaryote Protein Families

The eukaryotic protein data set was taken from Ku et al. (2015), hereafter referred to as the EK data set. It contains 21,146 protein sequences from 55 eukaryotic genomes from six different supergroups. The data set was divided into two different matrices: one for 1,060 protein families shared in photosynthetic eukaryotes and densely distributed in cyanobacteria (6,528 sequences, corresponding to block A, B, and C in Ku et al. 2015) and another for 1,397 protein families present in the eukaryotic common ancestor that are likely to correspond to the origin of the mitochondrion (14,618 sequences corresponding to block E in Ku et al. 2015).

For each data set, a presence/absence pattern was constructed. In the presence/absence patterns, each row corresponds to a species and each column to a protein family, binary elements of the matrix indicate presence or absence in the respective genome. Phylogenetic reference trees for the AR and EK data sets were taken from Nelson-Sathi et al. (2015) and Ku et al. (2015), respectively. For the PL data set, the reference tree was assembled from Schönfeld (2012) based on Bayesian inference of trees for the individual genes. Internal nodes are designated as HTUs (hypothetical taxonomic units), terminal nodes as OTUs (operational taxonomic units).

BLAST against Cyanobacterial Genomes

The 15,588 protein sequences in the PL data set were blasted against 94 cyanobacterial genomes retrieved from RefSeq, NCBI (version September 2016, listed in supplementary table 1, Supplementary Material online). Hits were filtered with a threshold of e-value ≤1E-10 and local identity ≥25%.

Calculation of Gain and Loss Events with Count

Version 10.04 of Count (Csűrös 2010), written in Java, was used. As input, Count requires a presence/absence pattern and the corresponding phylogenetic reference tree. Count's three methods for the analysis of gene evolution—two methods of maximum parsimony, Dollo (DP) and Wagner (WP) and the phylogenetic birth-and-death model (BD)—were tested. All models try to achieve an optimal score, with the exact optimization criterion varying according to the settings. Although using the same mathematical approach, the difference between the two parsimony methods here is, that DP weights a loss of a gene less than a gain while WP has the same cost for a loss or a gain of a gene. The reference tree and the appropriate presence/absence pattern were loaded into Count (branch lengths are ignored in parsimony models and were not used for the BD model). The settings were then optimized using likelihood, a necessary step in order to use the birth-and-death model. All other model parameters used were the default Count parameters (Groussin et al. 2016). The following settings were used: the model type was the gain–loss type, the family size distribution at the root was set to Poisson, lineage-specific variation was left unspecified, the gain variation across families was set to 1 for the edge length, the loss and the gain rate. The maximum number of optimization rounds was set to 100 with a convergence threshold on the likelihood of 0.1. The results of the different methods were displayed for each Count record in the graphical user interface, and then evaluated using a Perl script. The respective phylogenetic trees were processed and the results were recorded.

Trees were drawn with FigTree v1.4.2 (http://tree.bio.ed.ac.uk/software/figtree; last accessed April 2018) from the results provided by Count. The gain and loss events of the protein families for the respective method were summed and mapped for each corresponding node, respectively, in the phylogenetic tree. For the phylogenetic birth-and-death model the computed numbers for each protein family were rounded up (≥ 0.5) and down (< 0.5), respectively.

Results

Reproducing Count's Results for the Origin of Archaeal Protein Families

To reproduce the result of Groussin et al. (2016), we analyzed the subset of the AR data set (Nelson-Sathi et al. 2015) that they analyzed using the phylogenetic birth-and-death model of Count. A comparison (supplementary fig. 1, Supplementary Material online) shows that the number of gains calculated here using Count versus gains calculated using Count in Groussin et al. (2016) differed only very slightly and only for two archaeal groups (Thermococcales—58 vs 56—and Haloarchaea—219 vs 215). The reasons why Count produced very slight differences for six out of 568 gain events at the roots of the groups in our analysis versus the results of Groussin et al. (2016) are likely due to numerical error and the use of heuristic optimization algorithms.

More important is the circumstance that Count attributed no gains to the root of the archaeal tree in our analyses, nor did it do so in Groussin et al. (2016) but it does calculate more than one origin for different protein families. More than one origin for one protein family may occur when the distance between the species in the tree which have this protein is quite large, for example, when a specific protein is present in the species on both outer edges of the reference tree but is not present in the center nodes. Count, when instructed to handle binary data, which it was not designed to do (Csűrös 2010), then calculates that this protein originates at two or more different positions in the tree. Supplementary figure 1*b*, Supplementary Material online, shows the number of different origins per archaeal protein family calculated here for the AR data set. For 1,726 of the 2,264 archaeal protein families analyzed, Count calculated a single gain event, for 451 protein families two different origin events, for 87 families three different origins and for four of the protein families 6 different origins. For none of the protein families did Count calculate an origin at the root of the archaeal reference tree (Groussin et al. 2016).

Count Does Not Recover a Loss Only Process

To see whether Count can recover even an obvious process of massive gain followed by differential loss when misapplied to binary (presence/absence) data, we examined plastid genomes. It is generally accepted that plastids arose from cyanobacteria via endosymbiosis (Schwartz and Dayhoff 1978). It is also generally accepted that plastid genomes underwent reduction during evolution (Ohyama et al. 1986), that many genes were transferred to the nucleus during evolution and that many gene losses from cpDNA occurred in independent lineages (Martin et al. 1998, 2002). Figure 1 shows the presence/absence patterns for chloroplast encoded proteins in a sample of photosynthetic eukaryotes. A BLAST search against 94 cyanobacterial genomes (supplementary table 1, Supplementary Material online) shows that 95% of the sequences (highlighted in supplementary fig. 2, Supplementary Material online) have readily identifiable homologs in cyanobacteria. The tree is rooted with *Cyanophora*, but other roots, including the red lineage have been proposed (Rodríguez-Ezpeleta et al. 2005).

Regardless of whether we use the parsimony or the birth-and-death options of Count, the program only counts about half of the 254 protein families as being present in the plastid ancestor (figs. 2 and 3*a*). The other half of the (n.b.) plastid-encoded proteins are reconstructed by Count to have been acquired after the initial symbiotic event leading to the origin of the plastid, during plant evolution. That is, Count indicated that the primary endosymbiotic event involved acquisition of half a plastid followed by later aquisition of the other half via LGT events in independent lineages. In the birth-and-death model that Groussin et al. (2016) used, Count reports that 86 protein coding genes were acquired once and 36 protein coding genes were acquired twice in the process of lineage diversification during plastid evolution. That is, Count calculates that the 122 genes that were acquired from cyanobacteria after lineage divergence during plant evolution and then laterally transferred among eukaryotes. Count does not specify donor or recipient lineages. Another five protein coding genes were acquired three times during plastid evolution.

In the 112 years since Mereschkowsky (1905) suggested that plastids arose from cyanobacteria, no one has seriously proposed a stepwise acquisition of plastid genomes. Rather, plastid endosymbiosis operates via mass acquisition of genes at the cyanobacterial origin of the organelle, followed by gene loss and transfer to the nucleus (Martin and Herrmann 1998; Martin and Müller 1998; Timmis et al. 2004; Archibald 2015). The methods used by Groussin et al. (2016), however, deliver a result that clearly suggests "continuous" LGT into and among the members of the eukaryotic lineage in order to construct plastids "on the fly" in independent eukaryotic lineages. That is important because the central argument of Groussin et al. (2016) was that Count "*supports the continuous acquisition of genes over long periods in the evolution of Archaea.*" The suspicion is that their methodology is systematically biased toward the inference of continuous acquisition and does not recover expected events of periodic massive gains followed by gradual differential loss even when that is the true process. This raises serious concerns about the analysis by Groussin et al. (2016).

Figure 2 shows the gain events calculated by the three models plotted against the reference tree. Eleven is the maximum number of gains at an OTU for the BD model (also high for WP with 17 gains) at *Pyramimonas parkeae*, a model organism for early evolved Viridiplantae (Satjarak and Graham 2017). Wagner Parsimony places the highest number of gain events (nineteen) at *Nephroselmis olivacea*, which is considered a descendant of the earliest-diverging green algae (Turmel et al. 1999). It should be noted that all models place a considerable number of gain events at the common ancestor of Rhodophyta, Hacrobia, and SAR.

Wagner Parsimony predicts the largest number of different gain events for the same protein families (fig. 3*a*)—eight different origins for *ycf20*, a family of unknown function and nine for *cysT*, a sulfate transporter. The BD model predicts a

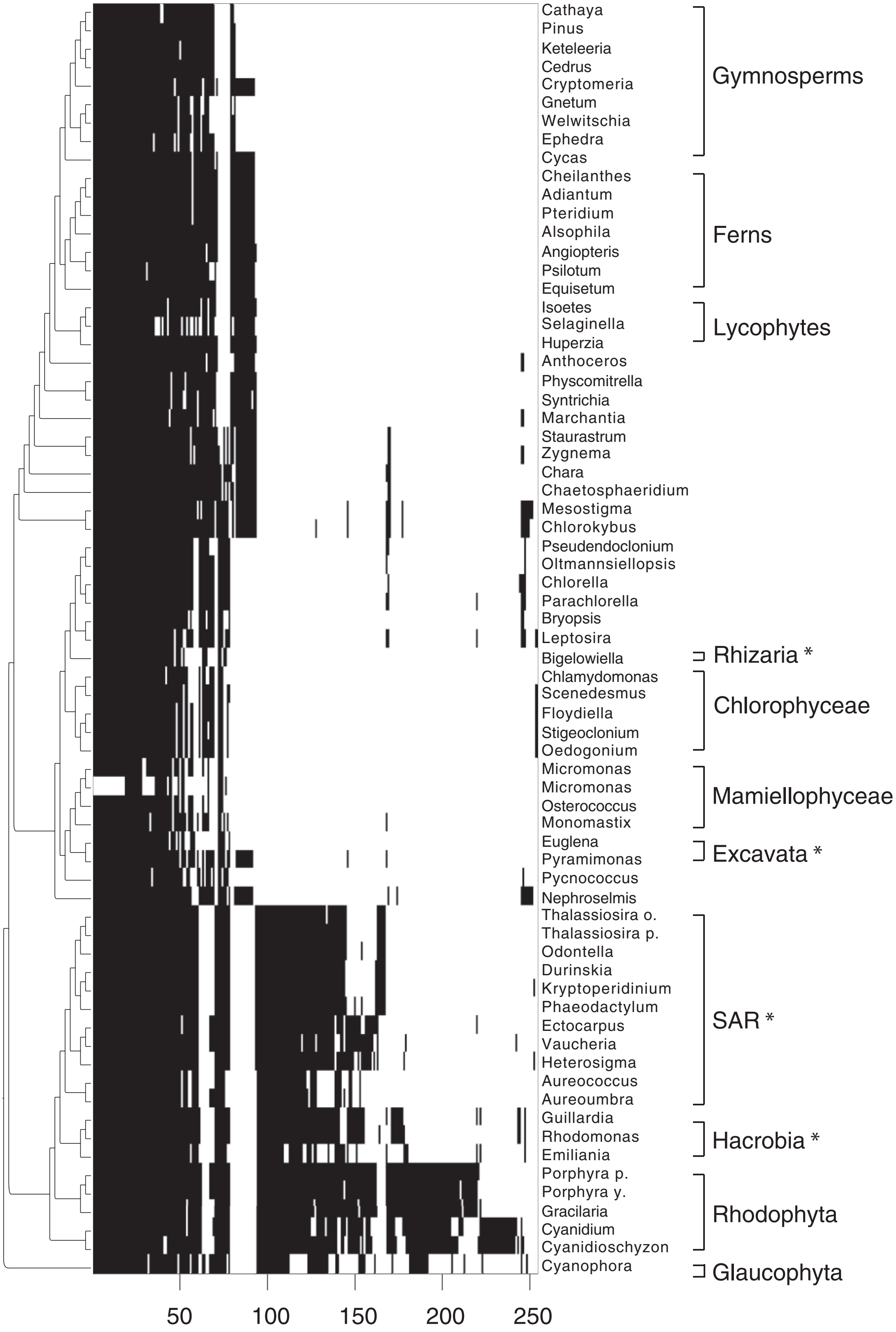

FIG. 1.—Presence/absence pattern of plastid protein families of the PL data set. Each black tick indicates the presence of a protein in an OTU. The number of protein families is indicated on the x axis. On the right side of the matrix are the OTUs, on the left the corresponding phylogenetic reference tree. Groups containing secondary plastids are marked with an *.

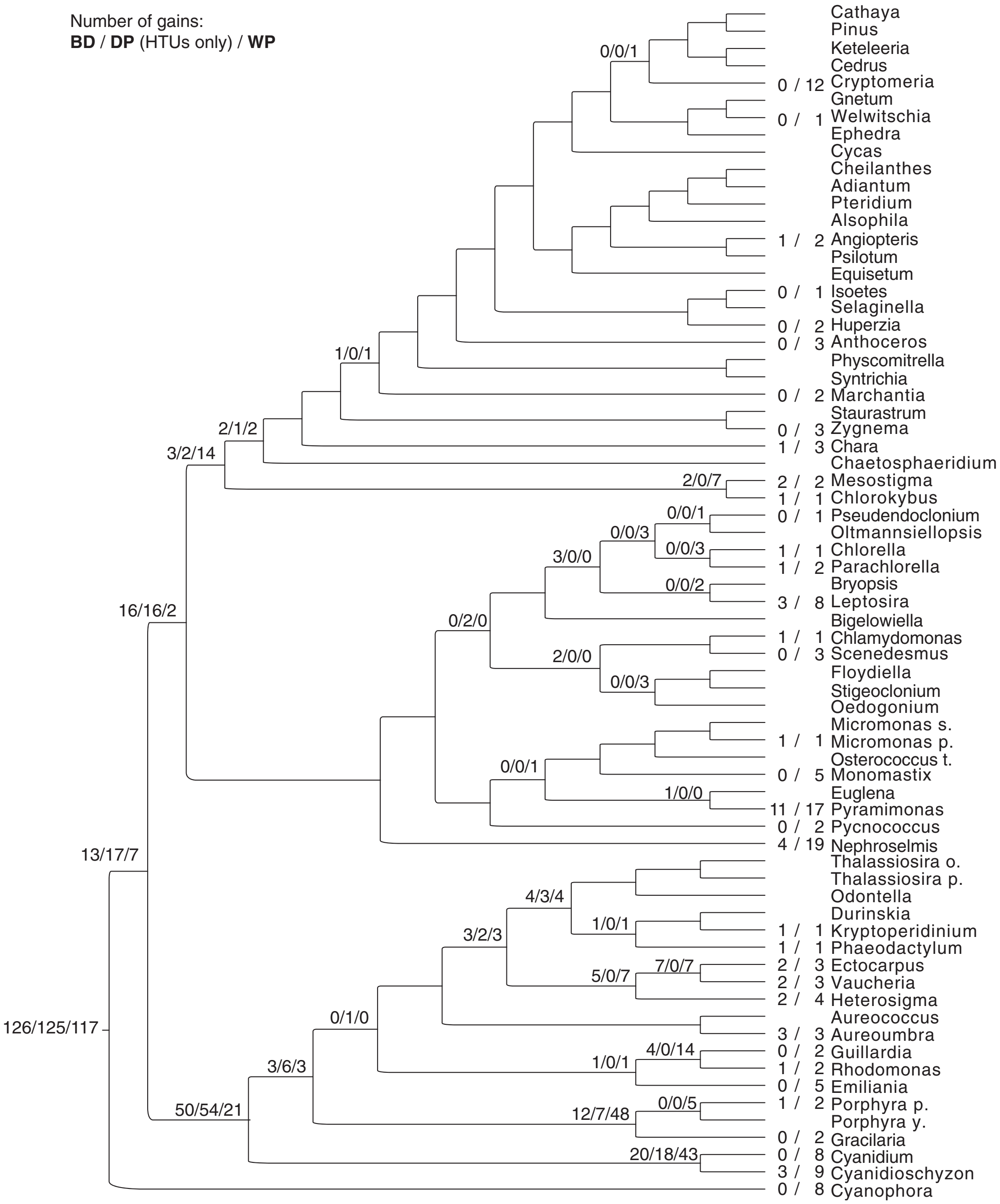

Fig. 2.—Phylogenetic reference tree for the PL data set with mapped gain events calculated with Count's traditional phylogenetic methods. Gain events for plastid protein families are depicted at the respective nodes in the following order, separated by slashes: Birth-and-Death model; Dollo Parsimony (only in the inner nodes); Wagner Parsimony. Inner and outer nodes where no values are plotted have no gain events according to the calculations of Count.

maximum of 4 different origins for *ycf47*, a poorly characterized probable protein exporter in thylakoid membranes. Dollo Parsimony does not predict more than one origin for any family, with only one gain event for all other 129 proteins occurring somewhere else throughout the tree—this is expected, as Dollo Parsimony favors the smallest number of different gain events. In other analyses (Martin et al. 2002) the corresponding patterns were identified as being the result of multiple independent gene losses. Both the BD and WP models predict a large number of gain events at the leaves of the reference tree—43 and 147, respectively (fig. 3*b*).

All three models in Count calculate at least one loss event per protein family for more than half of the families in the data set (supplementary fig. 3, Supplementary Material online). However, the number of gains (LGTs or convergent gene sequence homology origin) and losses per protein family is in

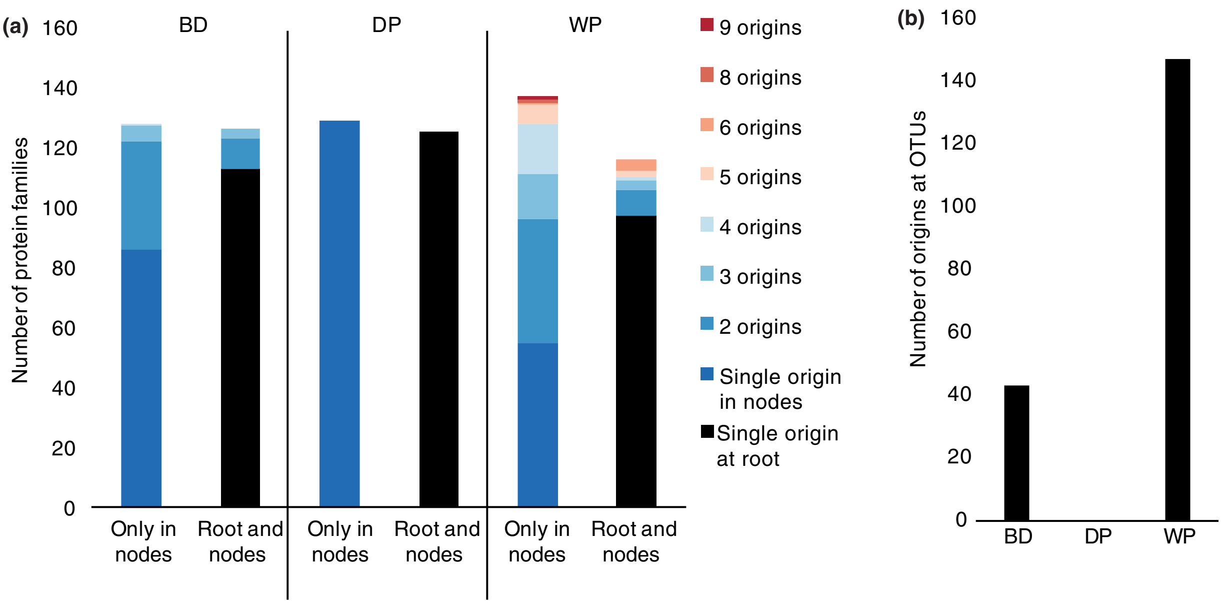

Fig. 3.—Multiple origins for the same protein families in the PL data set calculated by Count. (*a*) Number of different gains per protein family (split by gains only in nodes or at the root and nodes) for each phylogenetic model in Count; single origins at the root are highlighted in black; a gradient from blue to red shows multiple origins for the same protein family. (*b*) Number of origins in the outer nodes of the tree for each phylogenetic model in Count.

the same order of magnitude. This is evident in the result of the functional annotation of gain and loss events done for the PL data set (fig. 4). We annotated 224 of the 254 families. With the exception of Dollo Parsimony for photosystem II proteins and Calvin cycle, the tree models in Count predict at least one gene gain event in all the functional categories.

The Birth-and-Death Model of Endosymbiosis Events

Current views of eukaryote origin have it that eukaryotes arose from a symbiotic association between an archaeal host lineage and a mitochondrial endosymbiont (McInerney et al. 2014; Martin and Tielens et al. 2017; Zaremba-Niedzwiedzka et al. 2017) involving gene transfers from endosymbiont to host (Timmis et al. 2004; Thiergart et al. 2012). The origin of plastids entailed an additional influx of genes at the origin of the plant lineage (Ku et al. 2015). Thus, mitochondria and plastids each are currently understood to have had different, single origins, where large portions of the endosymbiont genomes entered the eukaryotic lineage. We checked the ability of the gene family size model used within Count to recover the massive episodic gene acquisition events at the origin of eukaryotes and chloroplasts, using presence/absence patterns prepared from the EK data set (Ku et al. 2015) and ignoring (as in Groussin et al. 2016) the fact that Count (Csűrös 2010) was not designed to analyze binary data. The distribution of those families is shown in figure 5, which is reproduced with permission from Ku et al. (2015).

Indeed, Count's BD model placed 1,410 of all 2,972 origin events for Group E proteins on the terminal edges of the phylogenetic reference tree (fig. 6*a*). The largest number of different gain events in a single OTU—98—was calculated for *Amphimedon queenslandica*, a sponge species known as a model for studying the origin and early evolution of animals (Srivastava et al. 2010). At the inner nodes of the tree the gain events were distributed almost uniformly, with only eight of the 53 inner nodes receiving no gain events with Count.

Out of 1,397 eukaryotic protein families belonging to Group E (see fig. 5) Count calculated that only 172 had a single origin at the root and no other gains anywhere else on the tree (fig. 6*b*). An additional 168 mitochondrial families were present at the root, however with additional origins spread throughout the tree (between one and 5 different origins). For 885 of the Group E protein families Count calculated between two and eight independent gain events (from prokaryotes via LGT or via eukaryote–eukaryote LGT). Count places a massive number of gain events at the leaves—1,410—for the Group E protein families (fig. 6*c*). It is important to recall that for the 2,585 genes families present in eukaryotes and prokaryotes in the data set of Ku et al. (2015), 87% show evidence for a single origin at the root using maximum likelihood methods (Ku et al. 2015). By contrast, Count reports that eukaryotes have acquired 88% of their genes independently from prokaryotes, but *from the same prokaryotic* donor each time, because otherwise the gene trees would not reflect a single origin relative to prokaryotic homologues (Ku et al. 2015). Clearly, within binary data, Count does not model adequately mass acquisitions such as those incurred at endosymbiotic events that gave rise to organelles, the reason being that it was not designed to model binary data (Csűrös 2010).

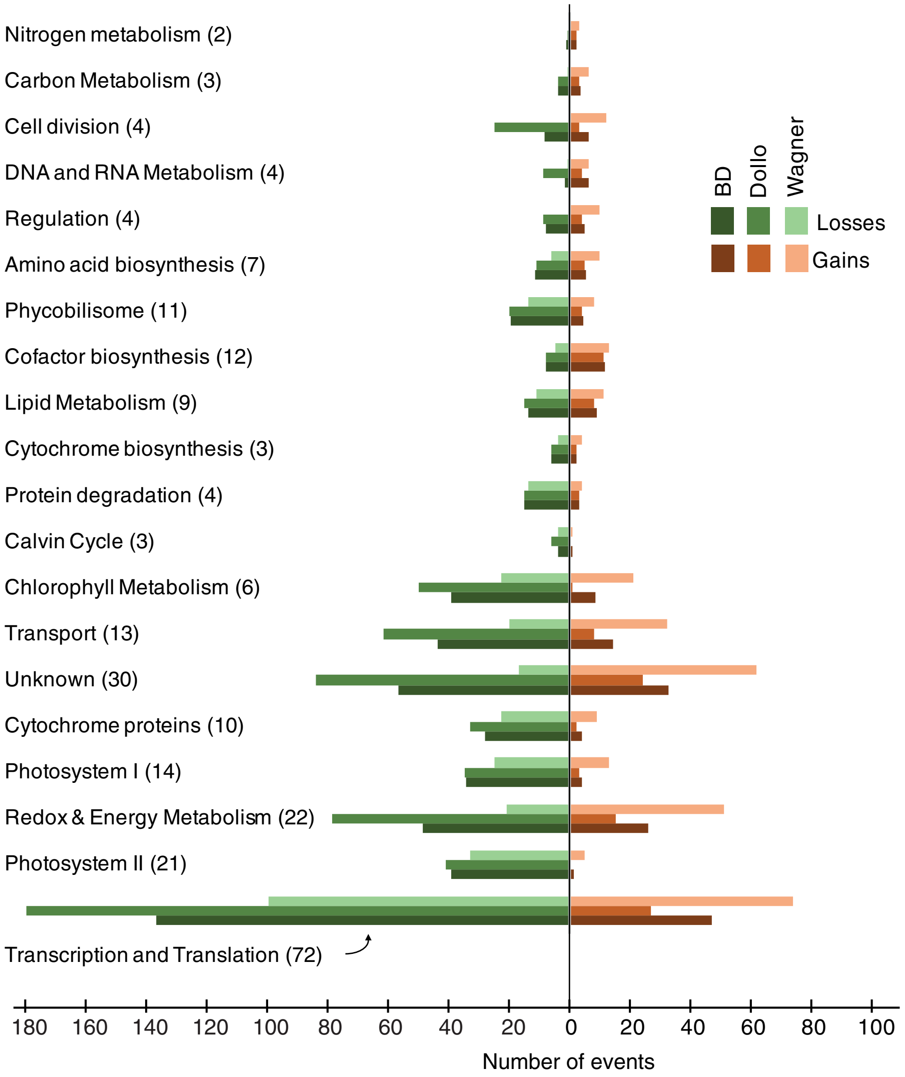

FIG. 4.—Gain and loss events for functional categories of protein families in the PL data set. The manual annotation resulted in 20 categories listed on the y axis, sorted by the prevalence in the presence/absence pattern (in parenthesis the total number of families in each category). Lost and gain events are shown on the left (greens) and right (oranges) side of the barplot, in the same scale, for the 3 different models in Count.

In the case of plastid families (Group A, B, and C in fig. 5), the genes for which are conspicuously widespread among cyanobacteria (fig. 5), Count produces the same effect: only 38 proteins out of 1,060 originate once and at the root of the subtree for plastid-containing species (fig. 6*a* and *b*). Count attributes another 191 families to the root and with additional origins elsewhere on the tree (between two and five different origins). According to Count, eukaryotes and plastids would have been acquiring the genes for the proteins that they need to survive "on the fly," that is via independent gains (of the same genes in independent lineages) during eukaryotic origin.

Furthermore, phylogenetic testing has shown that the vast majority of eukaryotic proteins in Group A, B, C, and E having homologues in prokaryotes are monophyletic, such that a single origin, not multiple origins, is the preferred model (Ku et al. 2015). Count does not recover that aspect of the data. Moreover, Ku et al. (2015) tested to see whether eukaryote to eukaryote LGT could account for the patchy distribution of the eukaryotic genes in figure 5. The result was that gene evolution in eukaryotes is resoundingly vertical (Ku et al. 2015), not lateral as in prokaryotes, hence the many independent origins (LGT) that Count infers do not reconcile with the phylogenies of the proteins underlying the presence/absence patterns with which Count operates. Out of the 1,761 calculated origins of the different plastid protein families, 339, almost a fifth, were found at leaves (fig. 6*c*). In only nine out of the 31 inner nodes of the plastid subtree there were no gain events of plastid families. Again, for the 2,585 gene families present in eukaryotes and prokaryotes in the data set of Ku et al. (2015), 87% show evidence for a single origin at the eukaryotic root using maximum likelihood methods (Ku et al. 2015).

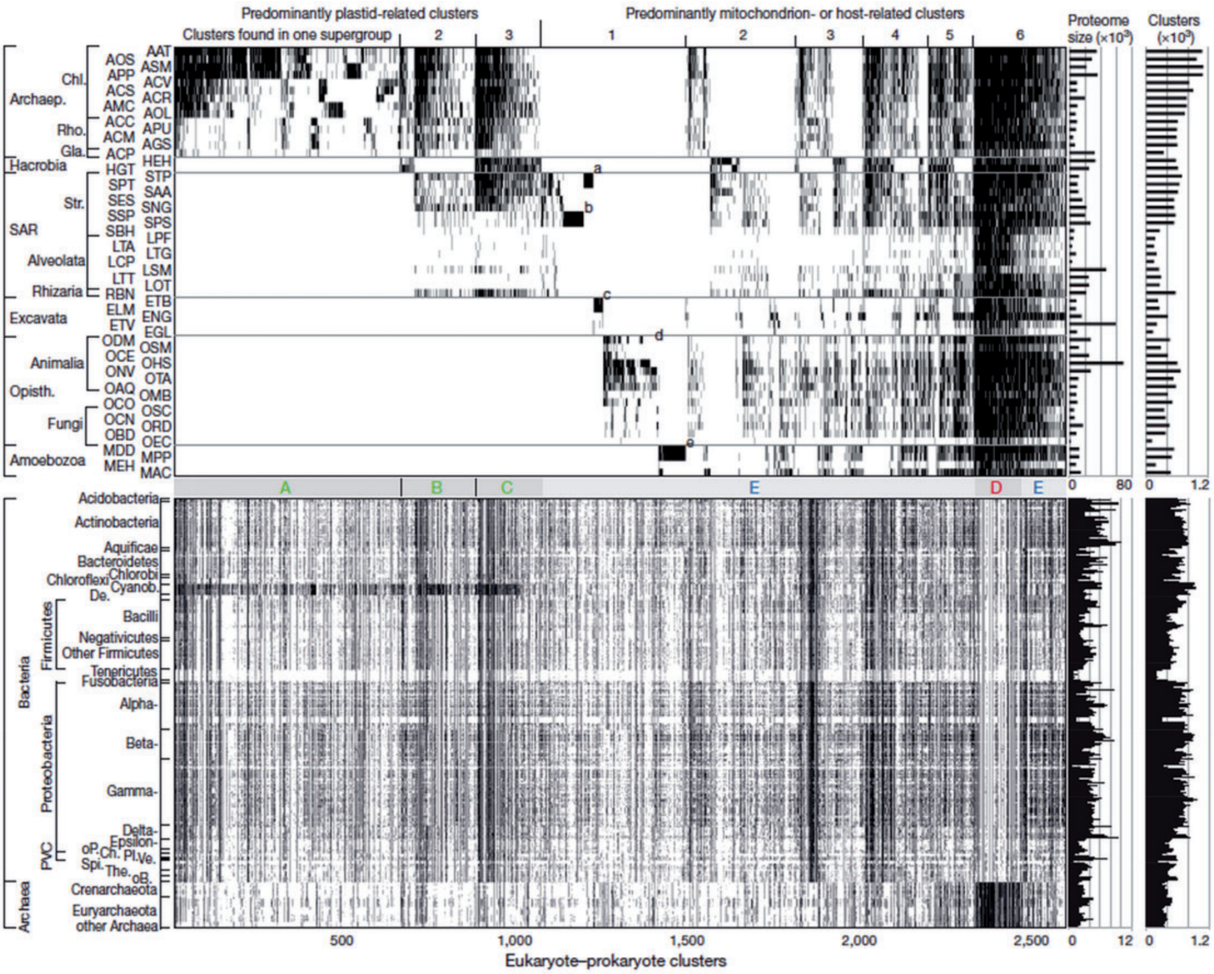

Fig. 5.—Gene distributions for eukaryotic genes. Reproduced with permission from Ku et al. (2015).

By contrast, Count reports that plastid bearing eukaryotes have acquired 96% of their genes independently from prokaryotes, but *from the same prokaryotic* donor each time, because otherwise the gene trees would not reflect a single origin relative to prokaryotic homologues (Ku et al. 2015). Clearly, Count is doing something very unusual with presence/absence pattern data in the case of mass acquisitions such as those incurred at endosymbiotic events that give rise to organelles. The same is almost certainly true for the mass acquisitions in archaea, where Count imposes a uniform process of acquisition upon the data, regardless of what the true process was.

Discussion

LGT is important in archaea (Wagner et al. 2017). Two recent studies have indicated that in archaea, gene acquisitions from bacteria can be episodic (Nelson-Sathi et al. 2012, 2015), similar results were found for transfers at the origin of eukaryotes and at the origin of plastids (Ku et al. 2015). Groussin et al. (2016) used a questionable application of Count (Csűrös 2010) as evidence that LGT in archaea is uniform, not episodic. We checked to see if the analysis carried out by Groussin et al. (2016) could recognize loss-only as the true model. We investigated proteins encoded in plastid genomes, which were sequestered from the cyanobacterial lineage ca. 1.6 Ba and have been vertically inherited in eukaryotes since, except during secondary endosymbiotic events. We analyzed the three different methods for ancestral reconstruction available in Count: the birth-and-death (BD) model, Dollo Parsimony (DP), and Wagner Parsimony (WP). The results obtained show that with BD and WP, Count distributes the origin of eukaryotic protein families uniformly throughout the tree and that more than one eukaryote LGT event is often calculated for the same protein family. With DP, there are also gain events throughout the tree, although not at the leaves (OTUs) and not twice for the same family.

The results of Count would suggest a process of continuous LGT for plastids and for eukaryotes, which runs counter to data (Ku et al. 2015; Ku and Martin 2016), the standard

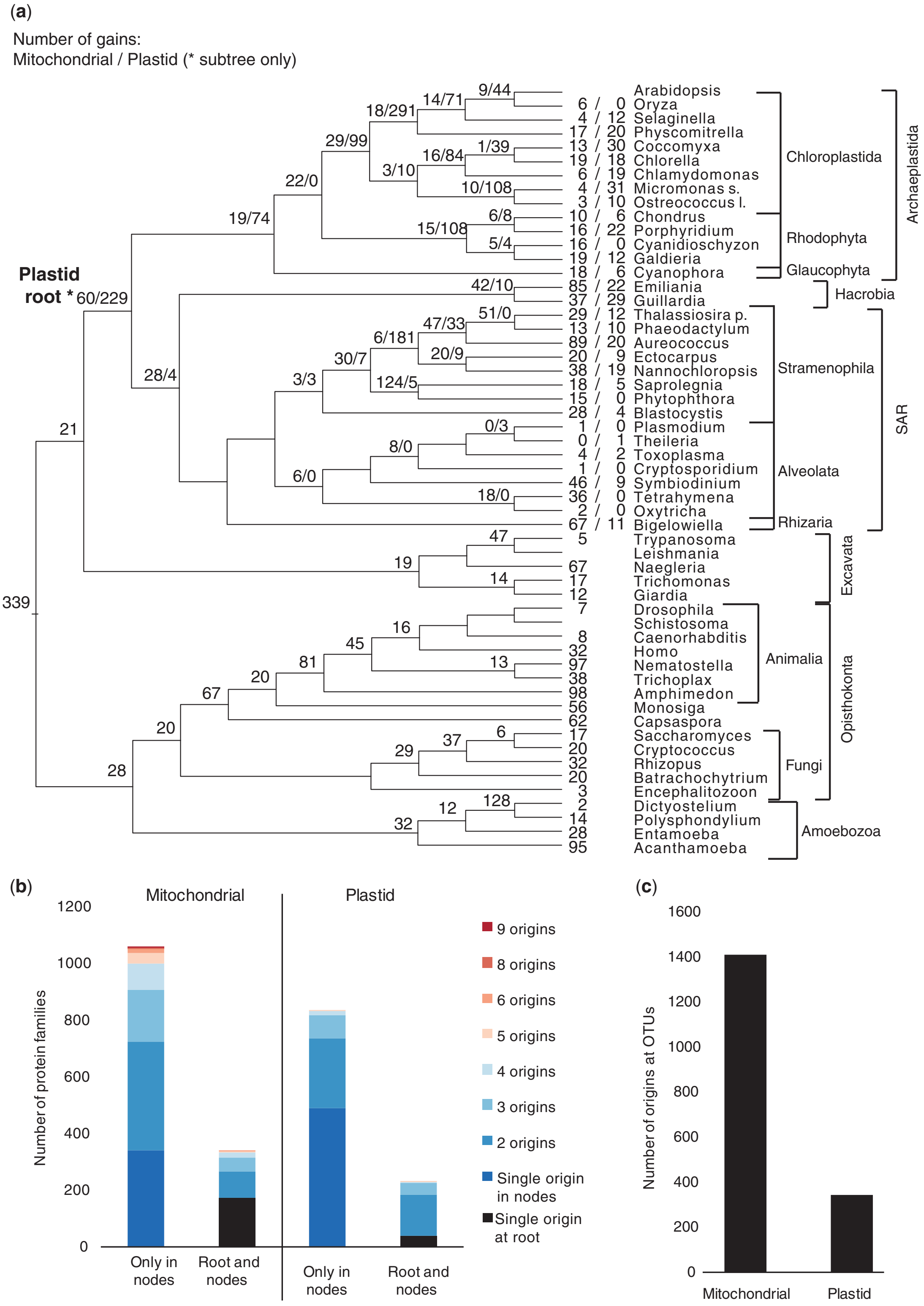

FIG. 6.—Gain events calculated by Count's birth-and-death model for mitochondrial and plastid protein families in the EK data set. (*a*) Reference tree for eukaryotes with mitochondrial and plastid origin events depicted at the respective nodes (separated by slashes), in this order. On the right, the 6 supergroups and the individual species (complete names in supplementary table 2, Supplementary Material online) are shown. The root for the plastid subtree is highlighted with a star (*). (*b*) Number of different gains per protein family (split by gains only in nodes or at the roots of each organelle's tree and nodes) for each phylogenetic model in Count; single origins at the root are highlighted in black; a gradient from blue to red shows multiple origins for the same protein family. (*c*) Number of origins in the outer nodes of the tree for each phylogenetic model in Count.

Darwinian paradigm of eukaryote evolution (Martin 2017), and eukaryote diploid genetics (Charlesworth et al. 2017). Count has it that different eukaryotic lineages independently assembled the collections of genes that make them eukaryotic (fig. 6) and that plastids independently assembled their genomes to look like reduced cyanobacterial genomes (figs. 2 and 3). Such inferences cannot be true.

We can see two potential explanations for why these analyses appear to fail. The first, and most obvious, is the misapplication of the models in Count to analyze binary data, rather than gene family size data. Groussin et al. (2016) carried out a form of data misrepresentation: gene families of any size in a genome are all coded as 1. A second, and more serious, problem with using Count to infer LGT events from binary data is the assumption of independence of gene families. Clearly this assumption is violated in the cases of acquisition and loss studied here (one large single gain followed by many independent losses).

However, the independence assumption could also distort inferences made in a more general setting (Lassalle 2017). There are two main, but related, effects. Firstly, the relative cost, to parsimony scores of likelihood, of LGTs are skewed. It becomes less costly, in terms of likelihoods, cheaper to posit separate LGTs for each gene family. Secondly, the independence of families means that the history for each gene family is inferred separately with no sharing of information across families. As each gene history is inferred using only the presence/absence pattern for that family (and the reference tree), the position of LGTs fit individually irrespective of whether they make sense in the larger context. The method that Nelson-Sathi et al. (2015) used to determine whether gene gain events correspond to the clade origins is independent of any reference tree and is even independent of gene presence/absence patterns: it compares two sets of trees with nonidentical leaf sets to determine whether their topologies were drawn from the same distribution.

In summary, the problem with the study by Groussin et al. (2016) is not that Count produces erroneous results. Count is a collection of different, efficient analysis methods for numerical characters, and is highly effective at reconstructing the evolution of gene family size. The problem with the study by Groussin et al. (2016) is that Count was applied to a problem that it was not designed to solve. The results show that using Count and especially these specific models was inappropriate for this data set. The simplifying choices made here, namely using binary presence/absence data and constant rates across families, seem to have a large influence and thus make the inference untrustworthy. The likelihood maximization used for the BD method requires a nonbinary distribution.

This is not a theoretical criticism: we have shown here that this problem has real and significant impact on inference. In particular, the systematic error explains the failure of Groussin et al. (2016) to recover the patterns of archaeal LGT uncovered in Nelson-Sathi et al. (2015) by methods that estimate the probability that two sets of trees were drawn from the same distribution.

The incorporation of dependence between gene families into methods like Count would be challenging both computationally and mathematically. Significant progress toward a heuristic solution has been made recently by Lassalle (2017). However, it could be still impossible to distinguish convincingly between different scenarios based only on presence/absence pattern data, there is simply insufficient information per gene family, and it might be statistically impossible to discriminate between radically different histories. The tests implemented by Nelson-Sathi et al. (2015) lacked the statistical power of full likelihood-based methods (Yang 2007), but on the other hand made no assumptions about the species tree and few assumptions about the process of LGT accumulation, gaining robustness in turn.

Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online.

Acknowledgments

This work was supported by grants from the European Research Council (666053) and the Volkswagen Foundation (93 046) to W.F.M., the German Israeli Foundation (I-1321-203.13/2015) to W.F.M. and E.H.-C., the Open University of Israel Research fund (504735) to E.H.-C., the Department of Science and Technology INSPIRE Faculty Award (DST/INSPIRE/04/2015/002935) to S.N.-S., and the New Zealand Bio-Protection Research Centre 2017 PI funding allocation to P.J.L.

Literature Cited

Albalat R, Cañestro C. 2016. Evolution by gene loss. Nat Rev Genet. 17(7):379–391.

Archibald JM. 2015. Endosymbiosis and eukaryotic cell evolution. Curr Biol. 25(19):R911–R921.

Charlesworth D, Barton NH, Charlesworth B. 2017. The sources of adaptive variation. Proc Biol Sci. 284(1855):20162864.

Csűrös M. 2010. Count: evolutionary analysis of phylogenetic profiles with parsimony and likelihood. Bioinformatics 26(15):1910–1912.

Dagan T, Martin W. 2007. Ancestral genome sizes specify the minimum rate of lateral gene transfer during prokaryote evolution. Proc Natl Acad Sci U S A. 104(3):870–875.

Daubin V, Moran NA, Ochman H. 2003. Phylogenetics and the cohesion of bacterial genomes. Science 301(5634):829–832.

Groussin M, et al. 2016. Gene acquisitions from bacteria at the origins of major archaeal clades are vastly overestimated. Mol Biol Evol. 33(2):305–310.

Ku C, et al. 2015. Endosymbiotic origin and differential loss of eukaryotic genes. Nature 524(7566):427–432.

Ku C, Martin WF. 2016. A natural barrier to lateral gene transfer from prokaryotes to eukaryotes revealed from genomes: the 70% rule. BMC Biol. 14(1):89.

Lassalle F. 2017. Ancestral genome estimation reveals the history of ecological diversification in *Agrobacterium*. Genome Biol Evol. 9(12): 3413–3431.

Martin WF. 2017. Too much eukaryote LGT. BioEssays 39(12):1700115.

Martin WF, et al. 1998. Gene transfer to the nucleus and the evolution of chloroplasts. Nature 1(4):276–165.

Martin WF, et al. 2002. Evolutionary analysis of Arabidopsis, cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus. Proc Natl Acad Sci U S A. 99(19):12246–12251.

Martin WF, Herrmann RG. 1998. Gene transfer from organelles to the nucleus: how much, what happens, and Why? Plant Physiol. 118(1):9–17.

Martin WF, Müller M. 1998. The hydrogen hypothesis for the first eukaryote. Nature 392(6671):37–41.

Martin WF, Roettger M, et al. 2017. Late mitochondrial origin is an artefact. Genome Biol Evol. 9(2):373–379.

Martin WF, Tielens AGM, Mentel M, Garg SG, Gould SB. 2017. The physiology of phagocytosis in the context of mitochondrial origin. Microbiol Mol Biol Rev. 81(3):e00008-17.

McInerney JO, O'Connell MJ, Pisani D. 2014. The hybrid nature of the Eukaryota and a consilient view of life on Earth. Nat Rev Microbiol. 12(6):449.

Mereschkowsky C. 1905. Über Natur und Ursprung der Chromatophoren im Pflanzenreiche. Biol Zent Bl 25:593–604.

Nelson-Sathi S, et al. 2012. Acquisition of 1,000 eubacterial genes physiologically transformed a methanogen at the origin of Haloarchaea. Proc Natl Acad Sci U S A. 109(50):20537–20542.

Nelson-Sathi S, et al. 2015. Origins of major archaeal clades correspond to gene acquisitions from bacteria. Nature 517(7532):77–80.

Ohyama K, et al. 1986. Chloroplast gene organization deduced from complete sequence of liverwort *Marchantia polymorpha* chloroplast DNA. Nature 322(6079):572–574.

Pisani D, Cotton JA, McInerney JO. 2007. Supertrees disentangle the chimerical origin of eukaryotic genomes. Mol Biol Evol. 24(8):1752–1760.

Popa O, Dagan T. 2011. Trends and barriers to lateral gene transfer in prokaryotes. Curr Opin Microbiol. 14(5):615–623.

Rodríguez-Ezpeleta N, et al. 2005. Monophyly of primary photosynthetic eukaryotes: green plants, red algae, and glaucophytes. Curr Biol. 15(14):1325–1330.

Satjarak A, Graham LE. 2017. Genome-wide analysis of carbohydrate-active enzymes in *Pyramimonas parkeae* (Prasinophyceae). J Phycol. 53(5):1072–1086.

Schönfeld B. 2012. The pattern and processes of genome change in endosymbionts old and new [doctoral dissertation]. New Zealand: Institute of Molecular BioSciences, Massey University.

Schwartz RM, Dayhoff MO. 1978. Origins of prokaryotes, eukaryotes, mitochondria, and chloroplasts. Science 199(4327):395–403.

Srivastava M, et al. 2010. The *Amphimedon queenslandica* genome and the evolution of animal complexity. Nature 466(7307):720–726.

Szöllősi GJ, Davín AA, Tannier E, Daubin V, Boussau B. 2015. Genome-scale phylogenetic analysis finds extensive gene transfer among fungi. Philos Trans R Soc Lond B 370(1678):20140335.

Szöllősi GJ, Tannier E, Daubin V, Boussau B. 2015. The inference of gene trees with species trees. Syst Biol. 64(1):e42–e62.

Thiergart T, Landan G, Schenk M, Dagan T, Martin WF. 2012. An evolutionary network of genes present in the eukaryote common ancestor polls genomes on eukaryotic and mitochondrial origin. Genome Biol Evol. 4(4):466–485.

Timmis JN, Ayliffe MA, Huang CY, Martin WF. 2004. Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes. Nat Rev Genet. 5(2):123–135.

Turmel M, Otis C, Lemieux C. 1999. The complete chloroplast DNA sequence of the green alga *Nephroselmis olivacea*: insights into the architecture of ancestral chloroplast genomes. Proc Natl Acad Sci U S A. 96(18):10248–10253.

Wagner A, et al. 2017. Mechanisms of gene flow in archaea. Nat Rev Microbiol. 15(8):492–501.

Williams TA, Foster PG, Cox CJ, Embley TM. 2013. An archaeal origin of eukaryotes supports only two primary domains of life. Nature 504(7479):231–236.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol. 24(8):1586–1591.

Zaremba-Niedzwiedzka K, et al. 2017. Asgard archaea illuminate the origin of eukaryotic cellular complexity. Nature 541(7637):353–358.

Associate editor: Eric Bapteste