Evidence That Plant-Like Genes in *Chlamydia* Species Reflect an Ancestral Relationship between Chlamydiaceae, Cyanobacteria, and the Chloroplast

Fiona S.L. Brinkman,[1,2,3,12,13] Jeffrey L. Blanchard,[4,5] Artem Cherkasov,[3] Yossef Av-Gay,[6] Robert C. Brunham,[7] Rachel C. Fernandez,[2] B. Brett Finlay,[2] Sarah P. Otto,[8] B.F. Francis Ouellette,[9] Patrick J. Keeling,[10] Ann M. Rose,[3] Robert E.W. Hancock,[2] and Steven J.M. Jones[11]

[1]Department of Molecular Biology and Biochemistry, Simon Fraser University, Burnaby, British Columbia, Canada, V5A 1S6; [2]Department of Microbiology and Immunology, University of British Columbia, Vancouver, British Columbia, Canada, V6T 1Z3; [3]Department of Medical Genetics, University of British Columbia, Vancouver, British Columbia, Canada, V6H 3N1; [4]Promega Corporation, Madison, Wisconsin 53711, USA; [5]National Center for Genome Resources, Santa Fe, New Mexico 87505, USA; [6]Department of Medicine, University of British Columbia, Vancouver, British Columbia, Canada, V5Z 4E3; [7]University of British Columbia Centre for Disease Control, Vancouver, British Columbia, Canada, V5Z 4R4; [8]Department of Zoology, University of British Columbia, Vancouver, British Columbia, Canada, V6T 1Z4; [9]Centre for Molecular Medicine and Therapeutics, Vancouver, British Columbia, Canada, V5Z 4H4; [10]Department of Botany, University of British Columbia, Vancouver, British Columbia, Canada, V6T 1Z4; and [11]Genome Sequence Centre, British Columbia Cancer Agency, Vancouver, British Columbia, Canada, V5Z 4E6

An unusually high proportion of proteins encoded in *Chlamydia* genomes are most similar to plant proteins, leading to proposals that a *Chlamydia* ancestor obtained genes from a plant or plant-like host organism by horizontal gene transfer. However, during an analysis of bacterial–eukaryotic protein similarities, we found that the vast majority of plant-like sequences in *Chlamydia* are most similar to plant proteins that are targeted to the chloroplast, an organelle derived from a cyanobacterium. We present further evidence suggesting that plant-like genes in *Chlamydia*, and other Chlamydiaceae, are likely a reflection of an unappreciated evolutionary relationship between the Chlamydiaceae and the cyanobacteria-chloroplast lineage. Further analyses of bacterial and eukaryotic genomes indicates the importance of evaluating organellar ancestry of eukaryotic proteins when identifying bacteria-eukaryote homologs or horizontal gene transfer and supports the proposal that Chlamydiaceae, which are obligate intracellular bacterial pathogens of animals, are not likely exchanging DNA with their hosts.

[Supplementary Material is available online at http://www.genome.org and at http://www.pathogenomics.bc.ca/BAE-watch.html.]

The Chlamydiaceae family of bacteria include several pathogens of animals and two important obligate human pathogens, *Chlamydia trachomatis* and *Chlamydophila pneumoniae* (Everett et al. 1999a; note that *Chlamydophila pneumoniae* was previously called *Chlamydia pneumoniae*). *C. trachomatis* is the causative agent of the sexually transmitted disease Chlamydia–the most frequently reported infectious disease in the U.S. and Canada and one of the leading causes of female infertility, ectopic pregnancy, and chronic pelvic pain (Division of STD Prevention 2000). It is also the causative agent of the ocular disease trachoma, one of the leading causes of blindness worldwide. *C. pneumoniae* causes acute respiratory infections and has been implicated in the development of atherosclerosis (Campbell et al. 1998). All Chlamydiaceae require intracellular infection of a host cell to replicate, complicating efforts to study these pathogens and develop a vaccine. To aid research, genome sequences have been obtained for five chlamydial strains comprising three species (Stephens et al. 1998; Kalman et al. 1999; Read et al. 2000; Shirai et al. 2000), and one of the most surprising observations from genome analyses has been the relatively high proportion of genes with highest similarity to plant sequences (Stephens et al. 1998). The obligate intracellular lifestyle of these bacteria has led to

[12]Corresponding author.
E-MAIL brinkman@sfu.ca; FAX (604) 291-5583.
[13]Present address: Department of Molecular Biology and Biochemistry, Simon Fraser University, 8888 University Drive, Burnaby, British Columbia, Canada, V5A 1S6.
Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.341802. Article published online before print in July 2002.

proposals that a *Chlamydia* ancestor obtained such genes from a plant or plant-like amoebal host organism by horizontal gene transfer (Stephens et al. 1998; Wolf et al. 1999b; Lange et al. 2000; Royo et al. 2000). Presumably, the intimate association between the Chlamydiaceae and their host cells would increase the chance of horizontal exchange of genes between host and bacterium. However, we present evidence that such plant-like genes in the Chlamydiaceae do not reflect horizontal gene transfer between these bacteria and their hosts. Rather, the plant genes appear to be derived from the cyanobacterial endosymbiont that gave rise to the chloroplast, and their similarity to homologs in the Chlamydiaceae reflects an ancient evolutionary relationship between Chlamydiaceae, cyanobacteria, and the chloroplast. Further analyses support our proposal that Chlamydiaceae are not likely exchanging DNA with their hosts and indicate the importance of evaluating the organellar ancestry of eukaryotic proteins.

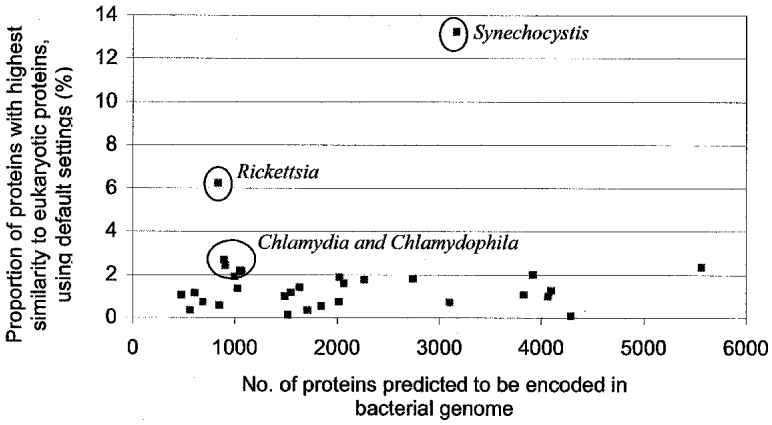

Figure 1 Proportion of proteins, predicted from complete bacterial genomes, which share highest similarity to eukaryotic proteins (according to analysis with default stringency settings; see http://www.pathogenomics.bc.ca/BAE-watch.html). Results for those organisms with a higher proportion than expected are circled. Similar results are obtained when different stringency cutoffs are used (see Supplementary Material available online http://www.genome.org).

RESULTS AND DISCUSSION

Analysis of Unusual Bacteria–Eukaryote Protein Similarities: Confirmation They Disproportionately Involve Chlamydiaceae, Cyanobacteria, and *Rickettsia*

We have developed an automated analysis of protein similarity based on BLAST (Altschul et al. 1997) to detect bacterial proteins notably more similar in primary sequence to eukaryotic proteins over other bacterial or archaeal proteins (and, conversely, eukaryotic proteins notably more similar to bacterial proteins over eukaryotic or archaeal proteins). A publicly available version of our analysis is at www.pathogenomics.bc.ca/BAE-watch.html (under the first three options). Although this analysis has obvious limitations (see Methods) and is not a substitute for phylogenetic analysis, we found it to be a useful aid in investigating bacteria–eukaryotic protein similarities at the primary sequence level.

This analysis showed that 65% of bacterial proteins identified with the highest similarity to a eukaryotic protein involved *Chlamydia*, *Chlamydophila*, *Synechocystis*, and *Rickettsia*, although these organisms only accounted for 14% of the genes analyzed (Fig. 1; Supplementary material; http://www.pathogenomics.bc.ca/BAE-watch.html). The proteins identified from *Rickettsia* were found to be disproportionately of the "energy production and conversion" functional category, and the *Synechocystis* and Chlamydiaceae proteins were found to be disproportionately similar to plant proteins. For *Rickettsia* and *Synechocystis* this was expected, due to the ancestral relationship between *Rickettsia* (an α-proteobacterium) and the energy-producing mitochondria and the ancestral relationship between *Synechocystis* (a cyanobacterium) and the chloroplast of plants and algae (Andersson et al. 1998; Reumann and Keegstra 1999). It is well known that a large proportion of organellar proteins are encoded by nuclear genes and that these proteins are targeted to the organelle posttranslationally using a transit peptide. It is thought that most of these genes were transferred from the endosymbiotic bacterium to the host nucleus during the transition of endosymbiont to organelle (Gray 1992). The "eukaryotic" genes identified from *Rickettsia* and *Synechocystis* are, therefore, not surprisingly predominantly similar to genes encoding proteins that function in the mitochondria and the chloroplast, respectively. A report proposing many horizontal gene transfer events between *Rickettsia* and eukaryote nuclear genes (Wolf et al. 1999b) did not include consideration of the movement of organellar genes into the nuclear genome, a phenomenon that has been known for some time (Weeden 1981) but is only now becoming more appreciated in eukaryotic genomics (Blanchard and Lynch 2000; Rujan and Martin 2001).

Plant-Like Genes in Chlamydiaceae: Plant Homologs Tend to Function in the Chloroplast

The notable number of plant-like genes in Chlamydiaceae genomes was more puzzling because Chlamydiaceae have no described relationship with any organelle. It was previously proposed that *Chlamydia* species obtained the genes from a host their ancestor had previously infected, such as a plant-like amoeba (due to the existence of *Chlamydia*-like organisms that infect *Acanthamoeba*, although *Acanthamoeba* is actually closely related to animals and fungi), whereas others suggested that they had simply obtained the genes from a plant (Stephens et al. 1998; Wolf et al. 1999b; Lange et al. 2000; Royo et al. 2000). However, analysis of multiple Chlamydiaceae genomes revealed a high level of conservation, suggesting they have been subjected to little horizontal gene transfer with other genera (Read et al. 2000). So where do the plant-like genes in Chlamydiaceae come from? Our comparison of eukaryotic genomes to those of Chlamydiaceae revealed that of the 18 cases of *Chlamydia* genes previously proposed to have been horizontally acquired from plants (Wolf et al. 1999a; Lange et al. 2000; Royo et al. 2000) 15 are similar to genes encoding proteins that function in the chloroplast in plants and the remaining 3 do not show a significant *Chlamydia*-plant relationship when subjected to phylogenetic analysis (Table 1). Furthermore, with the completion of the first plant genome (The Arabidopsis Genome Initiative 2000), we identified an additional 19 Chlamydiaceae proteins that

Table 1. Subcellular Localization in Plants of Proteins Similar to *Chlamydia* Proteins According to Low-Stringency BAE- (bacteria, archaea, and eukarya) Watch Analysis[a]

NCBI GI No.	Protein description	Subcellular localization in plants[b]
4377270	Glycyl tRNA synthetase	Chloroplast
4376626	ADP/ATP translocase[c]	Chloroplast
4376667	Glycogen hydrolase[c]	Chloroplast
4377189	GTP cyclohydratase and DHBP synthase	Chloroplast
4377237	Beta-ketoacyl-ACP synthase[c]	Chloroplast
4376686	Enoy-acyl-carrier reductase[c]	Chloroplast
4376591	Thioredoxin reductase[c]	Chloroplast
4377185	Metal transport P-type ATPase	Chloroplast
4377346	Similar to NA^+/H^+ antiporter	Chloroplast
4376981	Phosphate permease[c]	Chloroplast
4376650	GcpE protein	Chloroplast
4376637	Tyrosyl tRNA synthetase	Chloroplast
4377360	Malate dehydrogenase[c]	Chloroplast
4376763	GTP-binding protein	Chloroplast
4376911	ADP/ATP translocase[c]	Chloroplast
3329179	Phosphoglycerate Mutase	Chloroplast
4377281	Glycerol-3-Phosphate Acyltransferase[c]	Chloroplast
4376993	ABC Transporter ATPase	Chloroplast
4376509	Deoxyoctulonosic Acid Synthetase[d]	Chloroplast
4376872	Sugar Nucleotide Phosphorylase[e]	Chloroplast
4377368	Shikimate 5-dehydrogenase[c]	Chloroplast
4377054	Geranyl transferase	Chloroplast
3328465	1-Deoxyxylulose 5-phosphate reductoisomerase	Chloroplast
6578112	rRNA methyltransferase	Chloroplast
3329217	HSP60	Chloroplast
3328745	Phosphoribosylanthranilate isomerase[c]	Chloroplast
6578104	Aspartate aminotransferase[c]	Chloroplast[f]
4377328	Polyribonucleotide nucleotidyltransferase[c]	Chloroplast[f]
4377362	Putative D-amino acid dehyrogenase	Chloroplast[g]
4377331	Cytosine deaminase	Chloroplast?[h]
4376915	Lipoate-protein ligase A	Mitochondrial
4377272	Glycogen synthase	N/A[i]
4377065	Dihydropteroate synthase[c]	N/A[i]
4377239	Inorganic pyrophosphatase[c]	N/A[i]
4376904	Uridine 5′-monophosphate synthase	N/A[i]
4377173	UDP-glucose pyrophosphorylase[c]	N/A[i]
4376815	GutQ/Kpsf family sugar-phosphate isomerse	Mitochondrial?[j]

[a]Genes from *Chlamydia pneumoniae* CWL029 most similar to plant genes were identified using BAE-watch, with a step ratio of 1 and filtering any Chlamydial sequences (i.e., BAE-watch tertiary level) from the analysis. The same analysis was repeated with the other *Chlamydia* genomes to detect additional genes in this genus that are most similar to plant sequences. These nonstringent criteria identified all *Chlamydia* genes previously reported as having been horizontally acquired from plants. We propose that most of these plant proteins were originally of chloroplast origin and their similarity with *Chlamydia* sequences reflects *Chlamydia*'s ancestral relationship with the bacterial ancestor of the chloroplast.

[b]Chloroplast localization predicted by ChloroP (Emanuelsson et al. 1999), unless otherwise noted, and mitochondrial signal predicted by iPSORT (http://HypothesisCreator.net/iPSORT/)

[c]Identified as horizontal gene transfer by Wolf et al. (1999a)

[d]Identified as horizontal gene transfer by Royo et al. (2000)

[e]Identified as horizontal gene transfer by Lange et al. (2000)

[f]No prediction of a chloroplast transit peptide by ChloroP, however experimental evidence indicates that the protein is targeted to the chloroplast.

[g]Chloroplast localization predicted by iPSORT but not ChloroP.

[h]Incorrect start site appears to be predicted for protein. ChloroP predicts a transit peptide at start of sequence that shares similarity with all homologous proteins.

[i]Phylogenetic analysis indicates *Chlamydia* protein not most related to plant protein.

[j]Incorrect start site may be predicted (69 bp upstream) however this is uncertain so this protein represents the most interesting case in terms of potential gene transfer between bacteria and plants, or transfer between chloroplast and nuclear plant genomes that didn't involve targeting of the protein back to the chloroplast organelle.

are most similar to plant proteins, and 15 of these plant proteins are chloroplast targeted, 2 are predicted to be mitochondrial, and the remaining 2 do not bear out a significant Chlamydiaceae-plant relationship after phylogenetic analysis (Table 1). Additional Chlamydiaceae genes have also been previously noted to share highest similarity with proteins encoded in the chloroplast genome (Wolf et al. 1999b). It therefore appears that the vast majority of plant-like genes in Chla-

mydiaceae correspond to plant genes that are derived from, and function in, the chloroplast.

Evidence Chlamydiaceae, Cyanobacteria, and the Chloroplast Share an Ancient, Ancestral Relationship

With apparent links between Chlamydiaceae and chloroplast genes, we wondered whether Chlamydiaceae share a closer relationship with the chloroplast and cyanobacteria than is presently recognized. Previous phylogenetic analysis using small-subunit ribosomal RNA sequences did indeed suggest that *Synechocystis* and Chlamydiaceae form sister groups (Nelson et al. 2000) and this was confirmed through a bootstrapped analysis we performed with more cyanobacterial, Chlamydiaceae, and chloroplast sequences (data not shown). However, such analysis does not group these lineages with high confidence. This is most likely due to a significant divergence time between these lineages, which severely limits the phylogenetic information (informative sites) available, and also reduces the number of gene sequences that can be analyzed adequately. However, for analysis of such evolutionary relationships, it is becoming increasingly apparent that one should investigate multiple analyses and that such analyses should be carefully chosen for their appropriateness given the level of divergence being investigated. Character-based analyses of more slowly evolving molecular features is another approach (Qiu and Palmer 1999) that appears suitable in this case. Genomic characters, such as the presence or absence of signature sequences, introns, or genes in conserved operons, have been previously used to delineate a number of major groupings, including uniting certain charophycean green algae with plants (Baldauf et al. 1990; Manhart and Palmer 1990), grouping fungi and animals to the exclusion of plants and protists (Baldauf et al. 1996), and developing our picture of animal phylogeny (Boore et al. 1995). We therefore analyzed the ribosomal superoperon of 36 complete microbial genomes and 10 chloroplast genomes, investigating gene acquisition and loss from this operon as a slowly evolving character-based analysis. We identified several unique shared characters that unite Chlamydiaceae and *Synechocystis*/cyanobacteria exclusively and additional nonunique shared characters (Fig. 2). Another previously published slowly evolving character-based analysis of an unspliced group I intron in 23S rRNA also supports a link between Chlamydiaceae and the chloroplast lineage (Everett et al. 1999b). These results are also supported by analysis of the incomplete genome of the Cyanobacterium *Synechococcus* sp. strain WH8102 (preliminary sequence data obtained from the DOE Joint Genome Institute (JGI) at http://www.jgi.doe.gov/JGI_microbial/html), which shares the same unique and nonunique characters. Thus multiple genomes from the cyanobacterial and Chlamydiaceae lineages support this sisterhood. In addition, all 10 completely sequenced chloroplast genomes that we analyzed also share these characters (see Fig. 2 for a representative chloroplast analysis and see Methods for a list of the others). However, there has been additional gene loss from the chloroplast ribosomal superoperon (primarily through apparent transfers of genes to the plant nuclear genome; Fig. 2; data not shown). These observations, together with the existence of a higher than expected proportion of apparent chloroplast protein homologs in Chlamydiaceae genomes (and some weak phylogenetic analyses), appear to link Chlamydiaceae with the cyanobacterial/chloroplast lineage.

Genome Composition Analysis Suggests Chlamydiaceae Are Not Exchanging Genes with Their Hosts

In further support of the lack of horizontal gene transfer between Chlamydiaceae and their eukaryotic hosts, we also find that chlamydial genomes have been subjected to a low rate of recent DNA exchange with organisms of differing G+C ratios. The average G+C ratio for the genome of a particular microbial organism is often characteristic, with regions of DNA of unusual G+C ratios sometimes thought to reflect recent horizontal transfer of DNA from an organism with a differing G+C ratio. For Chlamydiaceae that are thought to infect only humans, the average G+C ratio of all genes or open reading frames (ORFs) from their genomes is 41% ± 2.5% (Table 2), whereas for humans the G+C ratio of their genes averages ~52% ± 8% (Nakamura et al. 2000; note that other mammals have a mean G+C ratio for genes that is similar to humans). Chlamydiaceae have a notably lower variance in their G+C ratio for genes than is observed for any other microbe whose genome has been sequenced to date (Table 2). In contrast, other bacteria, such as *Neisseria* species that have been shown to undergo frequent horizontal gene transfer, exhibit a much higher variance in %G+C for genes in their genomes (standard deviation up to ± 7%; Table 2). Although analysis of variance in gene %G+C for genomes cannot reveal horizontal acquisition of genes of the same G+C ratio and other factors such as level of gene expression can affect G+C ratios for a given gene, this low variance for whole chlamydial genomes is consistent with the lack of horizontal gene transfer suggested from the unrelated analysis of gene conservation and gene synteny in complete Chlamydiaceae genomes (Read et al. 2000). The apparently clonal nature of *Chlamydia* (and apparent lack of horizontal gene transfer) may be due to their ecological isolation from other bacteria, as a result of their intracellular lifestyle (Read et al. 2000).

Expanding the Analysis to Other Bacteria: Many Bacteria–Eukaryotic Protein Similarities May Reflect Bacterial Origin of Mitochondria and the Chloroplast

To further evaluate the involvement of organellar proteins in cases where bacterial genes are most similar to eukaryotic genes, we conducted a comparison of 162,003 genes from 37 bacterial and eukaryotic genome sequences (http://www.pathogenomics.bc.ca/BAE-watch.html). Although computational identification of organelle targeting signals has limitations (Emanuelsson et al. 2000), we found that the majority of bacterial proteins that are most similar to eukaryotic proteins share similarity to proteins that are known, or are proposed by `TargetP` analysis, to function in mitochondria or chloroplast organelles (see http://www.pathogenomics.bc.ca/BAE-watch.html and the section entitled "Bacterial proteins most similar to eukaryotic proteins"). Although *Chlamydia*, *Synechocystis*, and *Rickettsia* contain a far greater proportion of eukaryote-like genes than all other bacterial genomes analyzed (Fig. 1; Supplementary Material is available online at http://www.genome.org), this shows that one must be careful when examining proteins that share unusually high similarity between bacteria and eukaryotes to consider the possibility that a gene has organellar ancestry. In essence, it would appear that the bacterial origin of mitochondria and the chloroplast, coupled with the apparent horizontal transfer of genes from the organellar genome to the nuclear genome

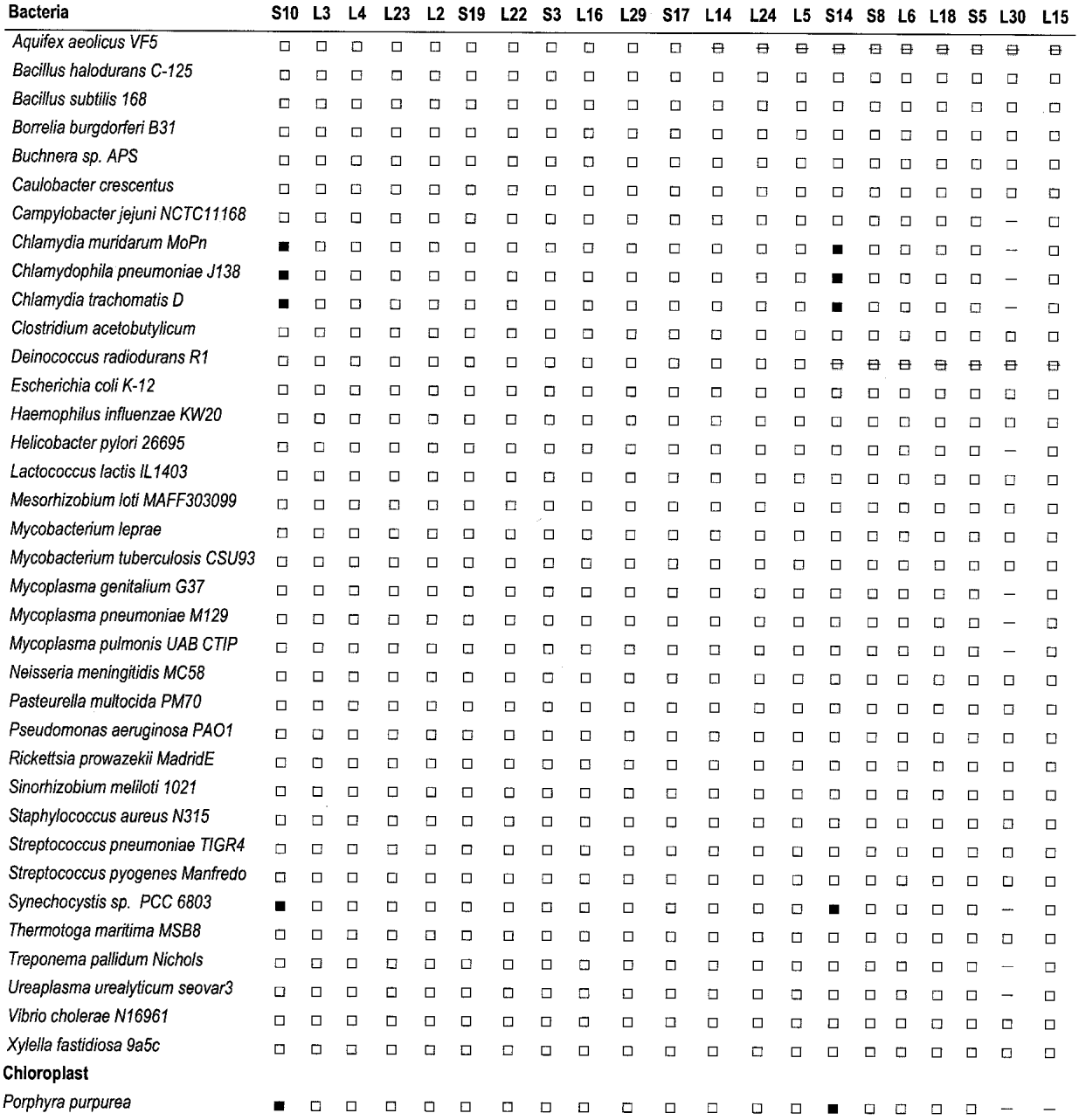

Figure 2 Unique shared-derived characters of the ribosomal super operon that unite cyanobacteria and Chlamydiaceae. Two unique shared-derived characters on the ribosomal super operon (the loss of ribosomal proteins S10 and S14) unite the Chlamydiaceae and cyanobacteria to the exclusion of other bacteria with genomes that have been completely sequenced (black boxes; note that S10 and S14 are present elsewhere on the chromosome). Loss of L30 (dashes; note that L30 does not appear to be present elsewhere in these genomes, according to TBLASTN analysis) is not a unique shared-derived character to the exclusion of all other bacteria but offers further support for a relationship between the Chlamydiaceae and cyanobacteria. In addition, all 10 chloroplast genomes examined (*Porphyra purpurea* chloroplast is shown as a representative) and an unfinished cyanobacterial genome (*Synechococcus* spp.) also share the same characters (i.e., loss of S10, S14, and L30 from the super operon); however, the chloroplasts are missing additional genes from this region (i.e., L15 in the region shown) that have been primarily transferred to the plant nucleus. Boxes with strikethroughs mark genes that have relocated in *Deinococcus* and *Aquifex* to form a separate operon. Note that the genome annotation for *Aquifex* did not report L29; however, we did positively identify this gene in *Aquifex* using TBLASTN. Another unique character uniting Chlamydiaceae, cyanobacteria, and the chloroplast, which is not illustrated in this figure, is that S10 is found as part of the separate S7/S12 operon in only the Chlamydiaceae, cyanobacteria, and chloroplast sequences examined.

of eukaryotes, must be considered a potential complicating factor of any analysis of bacterial–eukaryotic protein similarity.

Implications

Our analysis indicates that that the plant-like genes in Chlamydiaceae are most similar to plant genes with protein prod-

Table 2. Percent G + C Mean and Standard Deviations Determined from All Predicted Protein Coding Regions for Complete Genomes of Pathogenic Bacteria (as of April 2001)

Organism	Approximate host range—"Primary" disease	Intracellular?	Notes regarding clonality and evidence of horizontally transferred regions	No. of protein-coding ORFs	G + C for ORFs > 300 bp[a]	
					Mean	S.D.
Neisseria meningitidis MC58	humans—meningitis	extracellular	Nonclonal, demonstrated horizontal transfer with other species	2025	52.4	6.9
Neisseria meningitidis Z2491	humans—meningitis	extracellular	Nonclonal, demonstrated horizontal transfer with other species	2121	52.6	6.5
Xylella fastidiosa 9a5c	plants—citrus variegated chlorosis	extracellular	Evidence of phage-mediated horizontal gene transfer	2766	53.4	5.4
Escherichia coli O157:H7	warm-blooded animals, including humans—diarrhea	facultative intracellular	Compared with E coli K12, has higher % G + C S.D. and more predicted horizontally transferred regions	5283	51.0	5.3
Mycoplasma pneumoniae M129	humans—mycoplasmal pneumonia	extracellular		677	40.3	4.9
Vibrio cholerae N16961 chrom. 2 (of 2)	humans, zooplankton, other aquatic life—cholera	extracellular	More genes than chr. 1 that appear to have origins outside alpha-proteobacteria to which Vibrio belongs; proposed megaplasmid origin	1092	46.9	4.3
Treponema pallidum Nichols	humans—syphillis	extracellular		1031	51.4	4.2
Pseudomonas aeruginosa PAO1	humans, a range of other animals—variety of opportunistic mucosal infections	extracellular		5565	67.0	3.8
Ureaplasma urealyticum serovar3	humans—urethritis	extracellular		611	29.3	3.8
Vibrio cholerae N16961 chr. 1 (of 2)	humans, zooplankton, other aquatic life—cholera	extracellular		2736	48.1	3.7
Borrelia burgdorferi B31	humans, rodents, tick vector—Lyme disease	facultative intracellular		850	28.7	3.6
Campylobacter jejuni NCTC11168	humans, fowl, cattle, sheep, dogs, cats—gastroenteritis	extracellular	Noted for lack of insertion sequences or phage-associated sequences	1634	30.6	3.5
Mycoplasma genitalium G37	humans—urethritis (opportunistic)	extracellular		480	31.4	3.5
Pasteurella multocida PM70	range of animals—fowl cholera, cattle septicemia, pig rhinitis	extracellular		2014	40.8	3.3
Helicobacter pylori 266695	humans—peptic ulcers and gastritis	extracellular	Conserved relative to the other *H. pylori* genome	1553	39.4	3.4
Haemophilus influenzae Rd-KW20	humans—upper respiratory infection and meningitis	extracellular	Evidence of horizontal transfer between *Neisseria* and *Haemophilus,* but not in this genome sequence	1709	38.5	3.4
Helicobacter pylori J99	humans—peptic ulcers and gastritis	extracellular	Conserved relative to the other *H. pylori* genome	1491	39.3	3.3
Mycobacterium tuberculosis CSU93	humans—tuberculosis	facultative intracellular		3918	65.6	3.3
Rickettsia prowazekii MadridE	humans, other animals, lice vector—epidemic typhus	obligate intracellular	Highly clonal	834	30.1	3.3
Chlamydophila pneumoniae AR39	humans—chlamydial pneumonia	obligate intracellular	Highly clonal	997	41.1	2.6
Chlamydophila pneumoniae CWL029	humans—chlamydial pneumonia	obligate intracellular	Highly clonal	1052	41.1	2.6
Chlamydophila pneumoniae J138	humans—chlamydia pneumonia	obligate intracellular	Highly clonal	1070	41.1	2.6
Chlamydia trachomatis D	humans—chlamydia	obligate intracellular	Highly clonal	894	41.5	2.3
Chlamydia muridarum MoPn	humans—chlamydia	obligate intracellular	Highly clonal	818	40.8	2.3

[a]The calculation appears to be more accurate when ORFs < 300 bp are omitted from the analysis, as evident from a comparison of two highly similar *Chlamydia pneumoniae* genomes that suggests there are increased errors in gene prediction for genes > 300 bp in length.
This table is sorted by percent G + C standard deviation (S.D.) for predicted coding regions (ORFs) > 300 bp. Although the sample size is small, this standard deviation appears to correlate with the clonality of the microbe (two-tailed *P*-value > 0.005 for a nonparametric/Spearman correlation when condition is ranked). A similar analysis that is continually updated and includes nonpathogens, including archaea, is available at http://www.pathogenomics.bc.ca/IslandPath.html.

ucts that function in the chloroplast. We propose that the high proportion of plant-like genes in Chlamydiaceae is not due to horizontal gene transfer with a plant or related organism, but rather is a reflection of an ancient, ancestral relationship between the Chlamydiaceae and the cyanobacterial ancestor of the chloroplast. Regardless of the degree of relatedness between Chlamydiaceae and cyanobacteria, analysis of both Chlamydiaceae and other bacteria indicates that organellar ancestry must be considered in any case where a eukaryotic gene shares higher-than-expected similarity to bacterial homologs. One may wonder why Chlamydiaceae and other bacteria contain genes that share notable sequence similarity with organellar genes when there are species such as *Synechocystis* and *Rickettsia* that share an even closer relationship with the ancestors of organelles. First it must be emphasized that the number of such genes is far fewer than the number of organellar genes that share a highest similarity to cyanobacterial or rickettsial genes (Fig. 1). This is particularly notable for nonchlamydial bacteria if a high step ratio filter is used (see Methods for step ratio description) because BLAST is known for ordering sequences poorly in its output (Koski and Golding 2001) and such filtering aids in the removal of such BLAST ordering artifacts. It is also becoming increasingly apparent that gene loss plays a significant role in bacterial genome evolution (Mira et al. 2001; Salzberg et al. 2001). From this study, and others (Salzberg et al. 2001), it is clear that many cases of unusual bacteria–eukaryotic gene similarities are most likely a reflection of gene loss in a related lineage, coupled with our currently small taxonomic sampling of data at the genomic level. For example, *Synechocystis* may have lost a gene that is still present in Chlamydiaceae and the chloroplast, making the chlamydial gene appear most similar to the chloroplast counterpart in our analysis. Indeed, our analysis is currently only based on a single completed cyanobacterial genome, so it is quite possible that other cyanobacteria may still have orthologs of the gene (and when identified, this gene would be expected to be most similar to the chloroplast homolog). Consistent with this, most cases of plant-Chlamydiaceae gene similarity notably lack a *Synechocystis* homolog for comparison (or the homolog appears to be a paralog). These isolated cases (far fewer than the number of cases of *Synechocystis* genes resembling chloroplast genes) probably reflect gene loss in the *Synechocystis* lineage.

The apparent lack of horizontal gene transfer involving *Chlamydia*, both from their eukaryotic hosts (this paper) and from other bacterial genera (Read et al. 2000; this paper), suggests that *Chlamydia* may be a useful model for studies of gene evolutionary rates and for determining to what degree factors other than horizontal gene transfer can affect certain genomic properties. The observation of an evolutionary relationship between Chlamydia and cyanobacteria could have significance for *Chlamydia* research, as existing knowledge of cyanobacteria may stimulate new ways of thinking about the function and control of pathogenic *Chlamydia*.

METHODS

Protein/Gene Datasets and Phylogenetic Analysis

We analyzed complete published eukaryotic genomes (*Homo sapien*, *Arabidopsis thaliana*, *Drosophila melanogaster*, *Caenorhabditis elegans*, and *Saccharomyces cerevisiae*) for genes most similar to bacteria and, conversely, complete published bacterial genomes for genes most similar to eukaryotes (all pathogens are listed in the Supplementary Table [available online at http://www.genome.org], as well as *Synechocystis* sp. PCC6803, *Escherichia coli* K12, *Bacillus subtilis* 168, *Aquifex aeolicus* VF5, *Buchnera* sp. APS, *Bacillus halodurans*, *Lactococcus lactis* ssp. lactis IL1403, and *Thermotoga maritima* MSB8). For the human proteins, the ENSEMBL March 2001 dataset freeze was used (originally called version 8.0). For the genomic character analyses of the ribosomal superoperon, additional analysis were performed on chloroplast genes from *Porphyra*, *Cyanophora*, *Odontella*, *Plasmodium*, *Euglena*, *Marchantia*, *Rice*, *Tobacco*, *Chlorella*, and *Nephroselmis*. (See Acknowledgments for links to associated genome sequence publications and genome centers.

Phylogenetic analysis was performed using the neighbor-joining method of PHYLIP (http://evolution.genetics.washington.edu/phylip.html) for prealigned 16S rRNA genes from the Ribosomal Database Project II (http://rdp.cme.msu.edu/) for the following organisms: *Pyrococcus furiosus* (i.e., an archaeal sequence used to root the tree), *Thermotoga maritima*, *Aquifex pyrophilus*, *Bacillus subtilis*, *Chlamydophila pneumoniae*, *Chlamydophila psittaci*, *Chlamydia muridarum*, *Chlamydia trachomatis*, *Synechococcus* PCC6301, *Synechocystis* PCC6803, *Microcystis viridis*, *Escherichia coli*, *Caulobacter crescentus*, *Rickettsia prowazekii*, *Zea mays* (mitochondrial sequence), and chloroplast sequences from *Chlamydomonas reinhardtii*, *Klebsormidium flaccidum*, *Zea mays*, and *Nicotiana tabacum*.

Bacteria–Eukarya Protein Comparison Method

All complete bacterial and eukaryotic genomes mentioned above were compared using BLAST (Altschul et al. 1997) and MSPCRUNCH to a database of all proteins, including SWISS-PROT, TREMBL, and human proteins from the ENSEMBL March 2001 dataset. The results were placed in an ACEDB database (http://www.acedb.org) and related using TaxIDs to taxonomy information from the National Center for Biotechnology Information (NCBI's) Taxonomy database. The resulting database was queried for those proteins most similar to bacterial proteins over eukaryotic proteins (and those eukaryotic proteins most similar to bacterial proteins). This approach capitalizes on the significant evolutionary distance between the three Domains of life of bacteria, archaea, and eukarya and the presence in genetic databases of a number of completely sequenced genomes from all three domains (this increases the significance of a protein from one domain being more similar to a protein from another domain). A step ratio scoring system (see below) was developed to further filter the results and identify proteins that are substantially more similar to a protein from another domain of life over proteins from the same domain. This scoring system is necessary to filter from the analysis any proteins that are highly conserved in all organisms that BLAST scoring alone may identify as most similar to another domain's protein by chance. Previous analyses of proteins with highest similarity to proteins from other domains of life have suffered from failing to use a sufficiently stringent scoring system or not, ensuring that their scoring system is flexible enough to handle varying rates of gene evolution. This scoring system has normalized, flexible cutoffs. The database front end also facilitates filtering of various taxonomic groups of organisms from the analysis to identify, for example, bacterial genes conserved in a genera or family that share significant similarity to eukaryotic genes. Proteins that are annotated by SWISS-PROT as being encoded in an organelle, or containing an organelle transit peptide according to TargetP (Emanuelsson et al. 2000), are specifically highlighted in the database because the ancestor of mitochondria and the chloroplast is known to be bacterial; so organellar genes, or organellar genes that have moved to the nucleus, tend to be most similar to bacterial genes (Andersson et al. 1998; Reumann and Keegstra 1999; Rujan and Martin 2001). A publicly available version of our analysis that has been expanded to analyze all bacterial genomes and to make

all cross-domain comparisons between bacteria, archaea and eukarya is available at www.pathogenomics.bc.ca/BAEwatch.html. Note that there are obvious limitations to this analysis: It only detects primary sequence similarities detected by BLAST, it is not useful for identification of proteins highly conserved between all domains of life, its effectiveness is limited by the number of known genes in databases (although this will improve over time), and it is limited by the accuracy of organellar transit peptide prediction algorithms.

Score Calculation for the Step Ratio Used to Calculate the Significance of a Match

The following is performed for each case of cross-domain similarity detected (for example, a query bacterial protein is found by BLAST to have highest similarity to a eukaryotic protein). First, a given query protein (in the example, the bacterial protein) is compared to itself using BLAST to generate a "self-blast" bit score for its alignment to itself. This value is used to normalize all bit scores in the BLAST output (i.e., each bit score in the BLAST output is divided by this self-blast bit score). The difference between each normalized bit score as you go down the list of hits is calculated and then the maximum of these differences (the most significant "step" down in the blast scores) is identified for all hits until a hit is observed to a protein belonging to the same domain as the query protein (for example, bacterial). The ratio of this maximum difference over the max ratio is the step ratio (the max ratio is this normalized bit score for the alignment of the query protein [i.e., bacterial protein] with its top hit [i.e., eukaryotic protein]). A high step ratio score therefore reflects a substantial drop in bit score between the top-hit (i.e., eukaryote) sequence and the first same-domain (i.e., bacterial) sequence in the BLAST output list. A high step ratio score cutoff therefore selects against proteins that are highly conserved in all organisms (highly conserved protein would not have much of a drop in bit score between a top hit protein and other proteins in the BLAST output). This facilitates the removal of proteins that BLAST records as being most similar to a protein of another domain that are essentially artifacts of the inability of BLAST to order similarly related sequences in their correct order (Koski and Golding 2001). We have found a step ratio score cutoff of 10 removes the majority of such undesirable highly conserved proteins from the analysis. However, this value may be adjusted by the user and often a higher value is required to reduce false-positives.

ACKNOWLEDGMENTS

We thank all Pathogenomics Project members (www.pathogenomics.bc.ca/people.html) for comments and suggestions, Olof Emanuelsson (Stockholm) for assistance with large-scale use of TargetP before software licensing was available, and the many genome centers that published sequence data required for this analysis (see http://www.tigr.org/tdb/mdb/mdbcomplete.html, http://www.ncbi.nlm.nih.gov/PMGifs/Genomes/linksOrg.html, and http://www.ncbi.nlm.nih.gov:80/PMGifs/Genomes/euk_o.html). This work was funded by the Peter Wall Institute for Advanced Studies. J.L.B.'s research was supported in part by the Promega Postdoctoral Fellowship program under the guidance of Michael Slater. Bioinformatics applications mentioned in this paper can be accessed through the Pathogenomics Project Web site at http://www.pathogenomics.bc.ca.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

The Arabidopsis Genome Initiative. 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408:** 796–815.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **25:** 3389–3402.

Andersson, S.G., Zomorodipour, A., Andersson, J.O., Sicheritz-Ponten, T., Alsmark, U.C., Podowski, R.M., Naslund, A.K., Eriksson, A.S., Winkler, H.H., and Kurland, C.G. 1998. The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria. *Nature* **396:** 133–140.

Baldauf, S.L., Manhart, J.R., and Palmer, J.D. 1990. Different fates of the chloroplast *tufA* gene following its transfer to the nucleus in green algae. *Proc. Natl. Acad. Sci.* **87:** 5317–5321.

Baldauf, S.L., Palmer, J.D., and Doolittle, W.F. 1996. The root of the universal tree and the origin of eukaryotes based on elongation factor phylogeny. *Proc. Natl. Acad. Sci.* **93:** 7749–7754.

Blanchard, J.L. and Lynch, M. 2000. Organellar genes: Why do they end up in the nucleus? *Trends Genet.* **16:** 315–320.

Boore, J.L., Collins, T.M., Stanton, D., Daehler, L.L., and Brown, W.M. 1995. Deducing the pattern of arthropod phylogeny from mitochondrial DNA rearrangements. *Nature* **376:** 163–165.

Campbell, L.A., Kuo, C.C., and Grayston, J.T. 1998. *Chlamydia pneumoniae* and cardiovascular disease. *Emerg. Infect. Dis.* **4:** 571–579.

Division of STD Prevention. 2000. Sexually Transmitted Disease Surveillance 1999. Centers for Disease Control and Prevention, September 2000.

Emanuelsson, O., Nielsen, H., and von Heijne, G. 1999. ChloroP, a neural network-based method for predicting chloroplast transit peptides and their cleavage sites. *Protein Sci.* **8:** 978–984.

Emanuelsson, O., Nielsen, H., Brunak, S., and von Heijne, G. 2000. Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *J. Mol. Biol.* **300:** 1005–1016.

Everett, K.D., Bush, R.M., and Andersen, A.A. 1999a. Emended description of the order Chlamydiales, proposal of Parachlamydiaceae fam. nov. and Simkaniaceae fam. nov., each containing one monotypic genus, revised taxonomy of the family Chlamydiaceae, including a new genus and five new species, and standards for the identification of organisms. *Int. J. Syst. Bacteriol.* **49:** 415–440.

Everett, K.D., Kahane, S., Bush, R.M., and Friedman, M.G. 1999b. An unspliced group I intron in 23S rRNA links Chlamydiales, chloroplasts, and mitochondria. *J. Bacteriol.* **181:** 4734–4740.

Gray, M.W. 1992. The endosymbiont hypothesis revisited. *Int. Rev. Cytol.* **141:** 233–357.

Kalman, S., Mitchell, W., Marathe, R., Lammel, C., Fan, J., Hyman, R.W., Olinger, L., Grimwood, J., Davis, R.W., and Stephens, R.S. 1999. Comparative genomes of *Chlamydia pneumoniae* and *C. trachomatis*. *Nat. Genet.* **21:** 385–389.

Koski, L.B. and Golding, G.B. 2001. The closest BLAST hit is often not the nearest neighbor. *J. Mol. Evol.* **52:** 540–542.

Lange, B.M., Rujan, T., Martin, W., and Croteau, R. 2000. Isoprenoid biosynthesis: The evolution of two ancient and distinct pathways across genomes. *Proc. Natl. Acad. Sci.* **97:** 13172–13177.

Manhart, J.R. and Palmer, J.D. 1990. The gain of two chloroplast tRNA introns marks the green algal ancestors of land plants. *Nature* **345:** 268–270.

Mira, A., Ochman, H., and Moran, N.A. 2001. Deletional bias and the evolution of bacterial genomes. *Trends Genet.* **17:** 589–596.

Nakamura, Y., Gojobori, T., and Ikemura, T. 2000. Codon usage tabulated from international DNA sequence databases: Status for the year 2000. *Nucleic Acids Res.* **28:** 292.

Nelson, K.E., Paulsen, I.T., Heidelberg, J.F., and Fraser, C.M. 2000. Status of genome projects for nonpathogenic bacteria and archaea. *Nat. Biotechnol.* **18:** 1049–1054.

Qiu, Y.L. and Palmer, J.D. 1999. Phylogeny of early land plants: Insights from genes and genomes. *Trends Plant Sci.* **4:** 26–30.

Read, T.D., Brunham, R.C., Shen, C., Gill, S.R., Heidelberg, J.F., White, O., Hickey, E.K., Peterson, J., Utterback, T., Berry, K., et al. 2000. Genome sequences of *Chlamydia trachomatis* MoPn and *Chlamydia pneumoniae* AR39. *Nucleic Acids Res.* **28:** 1397–1406.

Reumann, S. and Keegstra, K. 1999. The endosymbiotic origin of the protein import machinery of chloroplastic envelope membranes. *Trends Plant Sci.* **4:** 302–307.

Royo, J., Gimez, E., and Hueros, G. 2000. CMP-KDO synthetase: A plant gene borrowed from Gram-negative eubacteria. *Trends Genet.* **16:** 432–433.

Rujan, T. and Martin, W. 2001. How many genes in *Arabidopsis*

come from cyanobacteria? An estimate from 386 protein phylogenies. *Trends Genet.* **17:** 113–120.
Salzberg, S.L., White, O., Peterson, J., and Eisen, J.A. 2001. Microbial genes in the human genome: Lateral transfer or gene loss? *Science* **292:** 1903–1906.
Shirai, M., Hirakawa, H., Kimoto, M., Tabuchi, M., Kishi, F., Ouchi, K., Shiba, T., Ishii, K., Hattori, M., Kuhara, S., et al. 2000. Comparison of whole genome sequences of *Chlamydia pneumoniae* J138 from Japan and CWL029 from USA. *Nucleic Acids Res.* **28:** 2311–2314.
Stephens, R.S., Kalman, S., Lammel, C., Fan, J., Marathe, R., Aravind, L., Mitchell, W., Olinger, L., Tatusov, R. L., Zhao, Q., et al. 1998. Genome sequence of an obligate intracellular pathogen of humans: *Chlamydia trachomatis*. *Science* **282:** 754–759.
Weeden, N.F. 1981. Genetic and biochemical implications of the endosymbiotic origin of the chloroplast. *J. Mol. Evol.* **17:** 133–139.
Wolf, Y.I., Aravind, L., Grishin, N.V., and Koonin, E.V. 1999a. Evolution of aminoacyl-tRNA synthetases—Analysis of unique domain architectures and phylogenetic trees reveals a complex history of horizontal gene transfer events. *Genome Res.* **9:** 689–710.
Wolf, Y.I., Aravind, L., Koonin, and E.V. 1999b. Rickettsiae and Chlamydiae: Evidence of horizontal gene transfer and gene exchange. *Trends Genet.* **15:** 173–175.

WEB SITE REFERENCES

http://evolution.genetics.washington.edu/phylip.html; PHYLIP home page.
http://HypothesisCreator.net/iPSORT/; iPSORT.
http://rdp.cme.msu.edu/; Ribosomal Database Project II.
http://www.acedb.org; ACEDB genome database system.
http://www.jgi.doe.gov/JGI_microbial/html; DOE Joint Genome Institute Microbial Genomics.
http://www.ncbi.nlm.nih.gov:80/PMGifs/Genomes/euk_o.html; NCBI's list of organelle sequences.
http://www.ncbi.nlm.nih.gov/PMGifs/Genomes/linksOrg.html; NCBI's list of genome centers.
http://www.pathogenomics.bc.ca/BAE-watch.html; BAE-watch database.
http://www.pathogenomics.bc.ca; BC Pathogenomics Project web site.
http://www.pathogenomics.bc.ca/IslandPath.html; IslandPath.
http://www.tigr.org/tdb/mdb/mdbcomplete.html; TIGR Microbial Database.

Received April 9, 2002; accepted in revised form May 23, 2002.