Analysis of Genome Content Evolution in PVC Bacterial Super-Phylum: Assessment of Candidate Genes Associated with Cellular Organization and Lifestyle

Olga K. Kamneva[1], Stormy J. Knight[1], David A. Liberles[1], and Naomi L. Ward[1,2,3,*]

[1]Department of Molecular Biology, University of Wyoming

[2]Department of Botany, University of Wyoming

[3]Program in Ecology, University of Wyoming

*Corresponding author: E-mail: nlward@uwyo.edu.

Accepted: November 29, 2012

Data deposition: All the sequence data used in this study are available via GenBank.

Abstract

The *Planctomycetes*, *Verrucomicrobia*, *Chlamydiae* (PVC) super-phylum contains bacteria with either complex cellular organization or simple cell structure; it also includes organisms of different lifestyles (pathogens, mutualists, commensal, and free-living). Genome content evolution of this group has not been studied in a systematic fashion, which would reveal genes underlying the emergence of PVC-specific phenotypes. Here, we analyzed the evolutionary dynamics of 26 PVC genomes and several outgroup species. We inferred HGT, duplications, and losses by reconciliation of 27,123 gene trees with the species phylogeny. We showed that genome expansion and contraction have driven evolution within *Planctomycetes* and *Chlamydiae*, respectively, and balanced each other in *Verrucomicrobia* and *Lentisphaerae*. We also found that for a large number of genes in PVC genomes the most similar sequences are present in *Acidobacteria*, suggesting past and/or current ecological interaction between organisms from these groups. We also found evidence of shared ancestry between carbohydrate degradation genes in the mucin-degrading human intestinal commensal *Akkermansia muciniphila* and sequences from *Acidobacteria* and *Bacteroidetes*, suggesting that glycoside hydrolases are transferred laterally between gut microbes and that the process of carbohydrate degradation is crucial for microbial survival within the human digestive system. Further, we identified a highly conserved genetic module preferentially present in compartmentalized PVC species and possibly associated with the complex cell plan in these organisms. This conserved machinery is likely to be membrane targeted and involved in electron transport, although its exact function is unknown. These genes represent good candidates for future functional studies.

Key words: genome evolution in PVC super-phylum, cellular compartmentalization, DUF1501 and *Planctomycetes*-specific cytochromes, mucin-degradation by *Akkermansia muciniphila*.

Introduction

The *Planctomycetes*, *Verrucomicrobia*, *Chlamydiae* (PVC) super-phylum is a group of six bacterial phyla: *Planctomycetes*, *Verrucomicrobia*, *Chlamydiae*, *Lentisphaerae*, *Poribacteria*, and OP3 (Wagner and Horn 2006). Some bacteria within this group exhibit complex eukaryote-like cellular compartmentalization, whereas others feature simple cell organization. All *Chlamydiae* have the simple cell structure common among Gram-negative bacteria. The cellular structure of OP3 species and a number of *Verrucomicrobia* and *Lentisphaerae* organisms, with regard to the presence or absence of intracellular membranes, remains unknown. However, all characterized *Planctomycetes* (Fuerst 2005), several *Verrucomicrobia* (Lee et al. 2009), one lentisphaerae species (Fuerst JA, personal communication), and the poorly characterized poribacteria (Fieseler et al. 2004) have a common cell plan that features an additional intracellular membrane and is unique to these bacteria. *Planctomycetes* also exhibit variations upon this common plan, featuring additional membrane-enclosed compartments of known or undetermined function (Fuerst 2005). These properties make PVC an attractive group for studying the evolution of biological complexity in bacteria.

Additionally, the PVC group provides an opportunity to study the evolution of bacterial lifestyles. There are four main types of lifestyle in bacteria and all of them can be found within the PVC super-phylum: free-living organisms lacking interaction with a eukaryotic host; mutualists—interaction occurs and both host and symbiont benefit from the interaction; commensals—interaction occurs but only the symbiont benefits; pathogens—interaction occurs and symbiont benefits from the interaction, whereas it has negative consequences on the host. The lifestyle of PVC organisms range from free-living soil and aquatic *Planctomycetes*, *Verrucomicrobia*, and *Lentisphaerae*, through commensal and mutualistic microbes of *Verrucomicrobia* and *Lentisphaerae*, to obligate pathogens of the phylum *Chlamydiae* (Balows et al. 1992) (supplementary table S1, Supplementary Material online). Phylum *Chlamydiae* includes intracellular pathogens causing diseases in human and animals. Various amoebae are recognized as a natural reservoir (Horn 2008) for chlamydial species, although in some cases the nature of the ecological relationship between amoebae and *Chlamydiae* is unclear. It has been shown that the presence of certain chlamydial symbionts increases the cytopathic effects of amoebae in their hosts, suggesting mutualistic interactions (Fritsche et al. 1998). In other cases, the presence of *Chlamydiae* has no effect on the amoebae, *Chlamydiae* do not become established in the amoebae, or chlamydial presence leads to death of the amoebae, which suggests commensal, transient, or pathogenic interactions, respectively (Gautom and Fritsche 1995). Phylum *Verrucomicrobia* contains the human intestinal commensal species *Akkermansia muciniphila*. Mechanisms underlying the *A. muciniphila*–human interaction are poorly understood (Derrien et al. 2011). Therefore, elucidation of lineage-specific evolutionary processes associated with the different PVC lineages could shed light on the emergence of pathogens and beneficial microbes, as well as the genetic basis of complex cell organization in PVC bacteria.

Specific organismal properties are often determined by the effects of small-scale evolutionary events (nucleotide/amino acid replacement and indel substitutions) on gene functions (Gu 1999; Knudsen and Miyamoto 2001; Britten 2002; Davids et al. 2002; Anisimova and Liberles 2007; Kamneva et al. 2010). However, the gene content of bacterial genomes can vary significantly even between closely related bacterial species, in terms of both the presence/absence and the size of particular gene families (Lefébure and Stanhope 2007). These larger scale differences can be attributed to horizontal gene transfer, gene duplication, emergence of novel protein-coding sequences from non-coding genomic regions, and subsequent loss of acquired genes due to either effects on fitness or stochastic processes. These genome content changes contribute to the emergence of different ecological and physiological properties in microorganisms (Ochman et al. 2000; Mira et al. 2001). Within the PVC super-phylum, acquisition of "membrane coat-like" proteins on several lineages has been proposed as a key evolutionary event leading to the complex cell plan in PVC bacteria (Santarella-Mellwig et al. 2010). Therefore, reconstruction of genome content evolution is fundamental for understanding the natural history of bacterial species and to study evolution of new traits in microbial genomes.

Although genome information is currently available for 20 species of *Planctomycetes*, *Verrucomicrobia*, and *Lentisphaerae* and for multiple chlamydial species (five genera), a comprehensive analysis of PVC genome dynamics has not been reported. This prompted the study described later. A number of methods are currently available for studying genome-level evolutionary change. Some of them operate on phyletic patterns (presence/absence of genes across genomes), ignoring evolutionary relationships between genes in gene families (Csűös 2010). This is particularly problematic when analyzing genomes of distantly related organisms, because a high degree of divergence over long periods of evolutionary time can obscure many duplications, losses, and transfers, and they will not be reflected in the phyletic profile. The recently described AnGST program (David and Alm 2010) performs a parsimony-based gene-tree species-tree reconciliation which takes gene phylogeny into account. This tool can be applied to a genome-scale data set. It also incorporates evolutionary events relevant for bacterial genomes: gene duplication, loss, and HGT. The method implemented in AnGST is also somewhat robust to gene phylogeny reconstruction errors through bootstrap tree amalgamation, the procedure of resolving uncertainties by incorporating reconciliation into the tree-building process. The gene tree with the lowest reconciliation cost is chosen from a collection of trees consistent with the set of bipartitions present in the input gene trees (David and Alm 2010). One main disadvantage of this approach is that it assumes a tree-like relationship between genomes in the data set. The problems associated with this assumption are discussed in greater details in the Results and Discussion section, although it presents a background tree on which other events can be mapped. There are a few further disadvantages with use of AnGST: 1) It requires a bifurcating species tree, which can be hard to obtain for large sets of distantly related organisms. There is a gene-tree species-tree reconciliation tool which can be used with a multifurcating species tree (Berglund-Sonnhammer et al. 2006), but it treats only duplication and loss, but not horizontal transfer. 2) AnGST employs parsimony to evaluate the inferred gene history; therefore, it requires somewhat arbitrary choice of event penalties and ignores branch lengths of the species tree. However, the only method allowing inference of a detailed gene history through species-tree guided gene-tree reconstruction implemented in a reliable statistical framework does not incorporate HGT events, which are prominent features of bacterial evolution (Åkerborg et al. 2009). Considering our need to account for gene-tree topology for distantly related species, together with the relevance of HGT events for

bacterial evolution, we therefore decided to use AnGST to study the evolution of PVC genomes. We inferred rates and patterns of gene birth, duplication, loss, and lateral transfer in PVC super-phylum evolution, and report the results of this large-scale analysis of gene family dynamics and lineage-specific patterns. To obtain insights into the evolution of host association in PVC bacteria, we examined the taxonomic affiliation of sequences similar to genes involved in carbohydrate metabolism in the *A. muciniphila* genome, examining the topology of every glycoside hydrolase and glycoside transferase-specific gene family. We found that the sequences most closely related to *A. muciniphila* genes that are potentially involved in mucin-degradation are from the genomes of various *Bacteroidetes* and *Acidobacteria* species. We also discovered a highly conserved genetic module preferentially present in PVC species possessing intracellular membranes, and possibly associated with the complex cell plan. The key protein of the module contains a DUF1501 domain and appears to have emerged on the PVC lineage. The presence of signal peptides and twin-arginine motifs in the module's proteins, as well as putative cytochrome-specific signatures, suggests involvement with electron transport and membrane targeting, although the exact function/localization is unknown. These genes represent promising candidates for future functional studies, enabled by recent advances in genetic tool development for PVC organisms (Domman et al. 2011; Jogler et al. 2011).

Materials and Methods

Sequence Data

To assemble a genome data set, we collected 26 sequenced genomes for organisms of the PVC super-phylum (all available genomes of *Planctomycetes*, *Verrucomicrobia*, and *Lentisphaerae* and six genomes of species representing five different chlamydial genera). We also included 73 representatives of other bacterial phyla with completely sequenced genomes, to obtain information on the evolutionary origin of genes in PVC genomes. The final list of 99 analyzed species is shown in supplementary table S1, Supplementary Material online. Protein sequences for genomes of interest were downloaded from GenBank (Benson et al. 2005). Taxonomic and ecological information for every organism in the data set was extracted from the GOLD database (Pagani et al. 2012). Phylum-level taxonomic units were used as taxon identifiers (supplementary table S1, Supplementary Material online). We included selected bacteria distantly related to the PVC super-phylum, to represent as much sequence diversity as possible within the constraints of computational time. Included in the data set were 35 host-associated organisms (pathogens, mutualists, or commensals) as well as 64 free-living soil and aquatic bacteria, to represent organisms with different ecological associations.

Gene Families, Alignments, and Gene Trees

Gene families were identified using OrthoMCL 1.4 (Li et al. 2003), with an inflation value of 1.5 and the threshold for expectation value for BLAST (Altschul et al. 1990) search set at 1×10^{-4}. Protein sequences were aligned using MUSCLE v3.7 (Edgar 2004) with default settings. Phylogenetic trees were reconstructed using PhyML (Guindon and Gascuel 2003) implementing the WAG + I + GAMMA evolutionary model (Reeves 1992; Yang 1993; Whelan and Goldman 2001), collecting five bootstrap samples for the bootstrap amalgamation procedure linked to gene-tree species-tree reconciliation. The model was selected as the best model for 44 gene families, tested independently (discussed later). It was applied to all 10,296 gene families, even those not specifically tested. Gene phylogeny for DUF1501-carrying proteins was reconstructed using PhyML and the WAG + I + GAMMA evolutionary model; however, it was not tested with bootstrap analysis. Domain architecture of genomic loci containing DUF1501-carrying proteins was visualized using the iTOL web-server (Letunic and Bork 2011).

Species Tree Reconstruction and Molecular Dating

A species tree topology was recovered for all 99 organisms in the data set. Forty-four gene families, represented by exactly one gene in every genome under consideration, were used as a set of phylogenetic markers. The evolutionary model of best fit was determined for every phylogenetic marker using ProtTest (Abascal et al. 2005). We found WAG + I + GAMMA to be the best-supported general model of evolution for all 44 groups. Phylogeny reconstruction was subsequently performed for every gene family using PhyML and the WAG + I + GAMMA evolutionary model, performing 20 bootstrap runs. The trees resulting from every bootstrap run, for every phylogenetic marker, were summarized using the consense program in the PHYLIP package (Felsenstein 1989) with default settings (extended majority rule). We found one fully resolved topology to be supported by the data set. This tree was rooted using the common ancestor of phyla *Firmicutes* and *Tenericutes* as an outgroup because this node of bacterial phylogeny was reported to be the most ancestral node in previous studies (Ciccarelli et al. 2006; David and Alm 2010) (supplementary fig. S1, Supplementary Material online).

Species tree topology was used to determine divergence times using PhyloBayes (3.3b) (Lartillot et al. 2009). We conducted two PhyloBayes runs with a CIR process model of rate correlation and five sets of temporal constraints that could be directly linked to fossil or geochemical evidence, or were previously reported in the literature (supplementary fig. S1 and table S2, Supplementary Material online). We collected 10,000 samples from each chain. Samples from every chain

were summarized with 2,000 sample points discarded as a burn-in and only including every 10th sample value to reduce autocorrelation between samples. Both chains led to similar results (supplementary fig. S2, Supplementary Material online). Final divergence time estimates were obtained from one chain following a burn-in of 2,000 cycles, after which trees were sampled every 10 cycles until the 10,000th cycle. Credible intervals of divergence time for every species tree node over the final chronogram are shown in supplementary figure S3, Supplementary Material online. We also used a concatenated alignment of 44 gene families represented by one gene in every genome in the data set, to test whether the data supported a phylogenetic network, using the Neighbour-Net algorithm implemented in the SplitsTree software package (Huson and Bryant 2006).

Ancestral Character Reconstruction

Ancestral state reconstruction for presence/absence of intracellular membranes in bacteria included in our data set was performed using Mesquite 2.75 (http://mesquiteproject.org, last accessed December 13, 2012), implementing the model of discrete character evolution with two parameters (character gain and loss rates) (Pagel 1999). We used the presence/absence of intracellular membranes as described in supplementary table S1, Supplementary Material online, and the species tree (supplementary fig. S1, Supplementary Material online). Mesquite ignores taxa for which the morphological character is not provided (species for which intracellular membrane presence/absence is unknown) and corresponding ancestral nodes. We inferred maximum likelihood estimates for gain and loss rates, the likelihood of the character distribution given the species tree and proportional likelihoods of both states (compartmentalized/uncompartmentalized cell) for every node of the species tree, except for those which lead to nodes with unknown cellular structure (supplementary fig. S10, Supplementary Material online).

Analysis of Genome Content Evolution

Genome content evolution was inferred using parsimonious gene-tree species-tree reconciliation as implemented in the AnGST program (David and Alm 2010). The most parsimonious reconciliation was inferred for a gene tree of every gene family containing two or more sequences, using event penalty values obtained as described in the section later. We did not enforce time-consistent reconciliation, allowing transfer to occur between any two lineages. The numbers of gain, loss, and transfer events were extracted from AnGST output for every gene family for every species tree lineage.

Genome Flux Analysis

We employed a method originally reported in (David and Alm 2010) to infer event penalties leading to minimal average difference in genome size between ancestor and descendant across the branches of the species tree (genome flux):

$$G_{\text{flux}} = \frac{\sum_{i=1}^{i=196} \sqrt{(\text{pGS}_i - \text{cGS}_i)^2}}{196};$$

where pGS_i and cGS_i denote ancestral (parent) and descendant (child) genome size for branch i, and 196 is the number of branches in the species tree.

We included all gene families containing two or more sequences in the genome flux analysis and used loss and speciation costs fixed at 1.0 and 0.0, respectively. Minimal genome flux was obtained with HGT penalty equal to 5, and duplication penalty equal to 3 (supplementary fig. S4, Supplementary Material online).

Analysis of Gene Loss, Duplication, and Transfer Rates

Although the most parsimonious gene-tree species-tree reconciliation provides direct assessment of a number of events (duplication, losses, and HGT) on every lineage of the species tree, it does not infer rates of those events. Here, we estimated rates of gene loss, duplication, and horizontal gene transfer across the lineages of the species tree in the following fashion:

$$\begin{aligned}\text{Rl}_i &= \text{Nl}_i/(L_i^{\times}\text{pGS}_i);\\ \text{Rd}_i &= \text{Nd}_i/(L_i^{\times}\text{pGS}_i);\\ \text{Rt}_i &= \text{Nt}_i/L_i;\\ \text{Rb}_i &= \text{Nb}_i/L_i;\end{aligned}$$

where Rl_i, Rd_i, Rt_i, and Rb_i, are the rate of losses, duplications, transfer, and gene birth events, respectively, occurred on the branch i; Nl_i, Nd_i, Nt_i, and Nb_i are the number of losses, duplications, transfer, and gene birth events, respectively, occurred on the branch i. L_i is the branch length of branch i in million years for transfer and birth rate and in thousand years for loss and duplication; pGS_i is the ancestral (parent) genome size for branch i.

Distribution of gene loss, duplication, and birth rates across the lineages of the species tree (after removing outliers and branches having rate equal to zero) and log transformation were modeled in a Bayesian framework as a normal distribution:

$$\text{Robs}_i|\mu,\tau \overset{\text{iid}}{\sim} \text{Norm}(\mu,\tau);$$

Using non-informative independence priors over mean (μ) and precision (τ):

$$\begin{aligned}\mu &\sim \text{Norm}(0.0, 1e-6);\\ \tau &\sim \gamma(0.001, 0.001);\end{aligned}$$

In the case of transfer rates, data (after removing outliers and branches having rates equal to zero and log

transformation) were modeled in a Bayesian framework as a mixture of two normal distributions:

$$\text{Robs}_i|\mu_1,\mu_2,\tau_1,\tau_2,\pi_1,\pi_2 \overset{\text{ind}}{\sim} \pi_1\text{Norm}(\mu_1,\tau_1), \\ +\pi_2\text{Norm}(\mu_2,\tau_2);$$

Using noninformative independence priors over means (μ_j), precisions (τ_j), and weights (π_j):

$$\mu_j \overset{\text{iid}}{\sim} \text{Norm}(0.0, 1e-6);$$
$$\tau_j \overset{\text{iid}}{\sim} \gamma(0.001, 0.001);$$
$$(\pi_1,\pi_2) \sim \text{Dirchlet}(1,1);$$

Predictive values were sampled from the posterior predictive distribution and compared with observed data; average differences between observed and predicted values were used to assess every model:

$\text{Rpred}_i|\mu,\tau \overset{\text{iid}}{\sim} \text{Norm}(\mu,\tau)$ for loss duplication and birth rates or

$\text{Rpred}_i|\mu_1,\mu_2,\tau_1,\tau_2,\pi_1,\pi_2 \overset{\text{ind}}{\sim} \pi_1\text{Norm}(\mu_1,\tau_1) + \pi_2\text{Norm}(\mu_2,\tau_2)$ for transfer rate;

$D_i = \text{Robs}_i - \text{Rpred}_i$

$Dav = \text{mean}(D_i)$

QQ plots were also examined to see whether observed and the predicted data were distributed in the same way.

The model was implemented in JAGS-3.2.0 using the rjags R package interface. Three chains of MCMC were run for 25,000 generations (for loss, duplication, transfer, and birth rates) sampling from the posterior distributions of the parameters: μ, τ, μ_j, τ_j, π_j Rpred_i, D_i, and Dav. Only every 5th sample was retained to reduce autocorrelation between samples and the initial 5,000 samples were discarded as a burn-in. Convergence of the chains was assessed using trace plots. Technical results of the modeling protocol are summarized in supplementary figures S5–S8, Supplementary Material online.

D_i values were also used to identify extreme data values. If 85% or more, D_i values for a particular i were smaller or larger than zero, we classified branch i as a branch with decelerated or accelerated rate of events (losses, duplications, transfers, and birth events).

Analysis of Donor Lineages in Transfer Events

We used a binomial test to detect statistical overrepresentation of various evolutionary lineages among those acting as donors in HGT events, assuming each transferred gene to be an independent event and all organisms to have equal ecological opportunity to transfer genes. The probability of an organism acting as an HGT donor was assumed to be proportional to genome size. We transformed the P values obtained for every donor lineage to generate heatmap charts. The transformation scales all P values to range from -1 to 1, where values approaching -1 correspond to underrepresented lineages, and values approaching 1 correspond to overrepresented lineages. Heatmap charts were generated using the gplots R package.

Phyletic Profiling

We employed phyletic profiling to search for gene families exhibiting profiles correlated with the presence/absence of intracellular membranes. We excluded PVC organisms with unknown cellular structures from this analysis (supplementary table S1, Supplementary Material online). We used Pearson correlation coefficients as a measure of similarity between membrane presence/absence and phyletic profile for every gene family, and retrieved protein groups with correlation coefficients above 0.8. We also used a BLASTP search with *Gemmata obscuriglobus* protein gi: 168700061 to characterize gene families with similarity to membrane coat-like proteins.

Analysis of Biological Functions

To link gene families to metabolic pathways, proteins (from one representative organism of every species with a KEGG-annotated genome) and their cellular pathway annotations were retrieved from the KEGG database. Gene families were assigned to pathways based on either existing KEGG annotation for a gene (for those genomes which were annotated in KEGG) or annotation of BLAST best hit for the genes of a gene family. If at least half of the genes from a gene family were assigned to one particular pathway, the entire protein family was assigned to this biological pathway. We employed a binomial test to determine statistical over- and underrepresentation of cellular pathways among the gene families affected by gene birth, duplication, transfer, and loss processes on certain evolutionary lineages. We transformed the P values obtained for every pathway to generate heatmap charts. The transformation scales all P values to range from -1 to 1, where values approaching -1 correspond to underrepresented pathways, and values approaching 1 correspond to overrepresented pathways.

The CAZY-based classification of carbohydrate metabolism-related genes from the *A. muciniphila* genome was downloaded from http://www.cazy.org (last accessed December 13, 2012).

Heatmap charts were generated using the gplots R package.

Results and Discussion

Species Tree

To characterize genome content evolution in the PVC super-phylum, we assembled a set of all available finished or draft genomes of organisms from phyla *Planctomycetes*, *Verrucomicrobia*, and *Lentisphaerae* (20 genomes) and six genomes of organisms belonging to phylum *Chlamydiae*.

We also included 73 species representing other bacterial phyla. We reconstructed a background "species tree" for the entire data set of 99 genomes using a set of phylogenetic markers present in every genome in one copy. The recovered species tree topology (fig. 1 and supplementary fig. S1, Supplementary Material online) was largely similar to species relationships reported in previous studies (Wagner and Horn 2006; Hou et al. 2008; Pilhofer et al. 2008; Kamneva et al. 2010). The presence of non-PVC species in the data set allowed us to identify a sister clade of the PVC super-phylum, containing phyla *Spirochaetes*, *Bacteroidetes*, and *Chlorobi* (fig. 1 and supplementary fig. S1, Supplementary Material online). This relationship was also recovered in previous studies conducted using a different set of species and phylogenetic markers (Hou et al. 2008).

Although we were able to recover a species tree, we also detected a strong signal for a phylogenetic network at the root of the tree (supplementary fig. S9, Supplementary Material online). This highlights the role of HGT in bacterial evolution, and supports the current view on evolution of living organisms which incorporates not only the signal of vertically inherited genetic information (unique ancestor/descendant relationship), but also the signal of lateral transfer events (several ancestors for an organism) (McInerney et al. 2011). Although significant empirical evidence of a network-like evolutionary history has accumulated for both pro- and eukaryotic organisms (Fitzpatrick et al. 2006; Sullivan et al. 2006; McDonald et al. 2008; Retchless and Lawrence 2010), the majority of methods evaluating gene history assume tree-like evolutionary relationships between genomes/organisms, as a background on which horizontal processes can occur. Several factors complicate inference of tree-like species relationships in bacteria. The vague definition of bacterial species arises from the fragmented nature of bacterial speciation and poorly defined species boundaries (Fraser et al. 2007; Retchless and Lawrence 2010). In the early stages of speciation, genetic isolation is normally established at genomic loci containing niche-specific genes, whereas recombination persists in many other regions of the genome (including those encoding genes that are highly connected in the genetic network and resistant to HGT, such as information processing genes) (Retchless and Lawrence 2007). However, after genetic isolation has been established for major parts of the genome, illegitimate recombination, or transfer events facilitated by various mechanisms may still be a source of genetic material from distantly or closely related organisms. At this stage, genes with low connectivity in the genetic network such as those encoding antibiotic resistance, photosynthesis, or other metabolism-related genes are transferred (Rivera et al. 1998; Cohen et al. 2010). Multiple genetic markers are often used to analyze

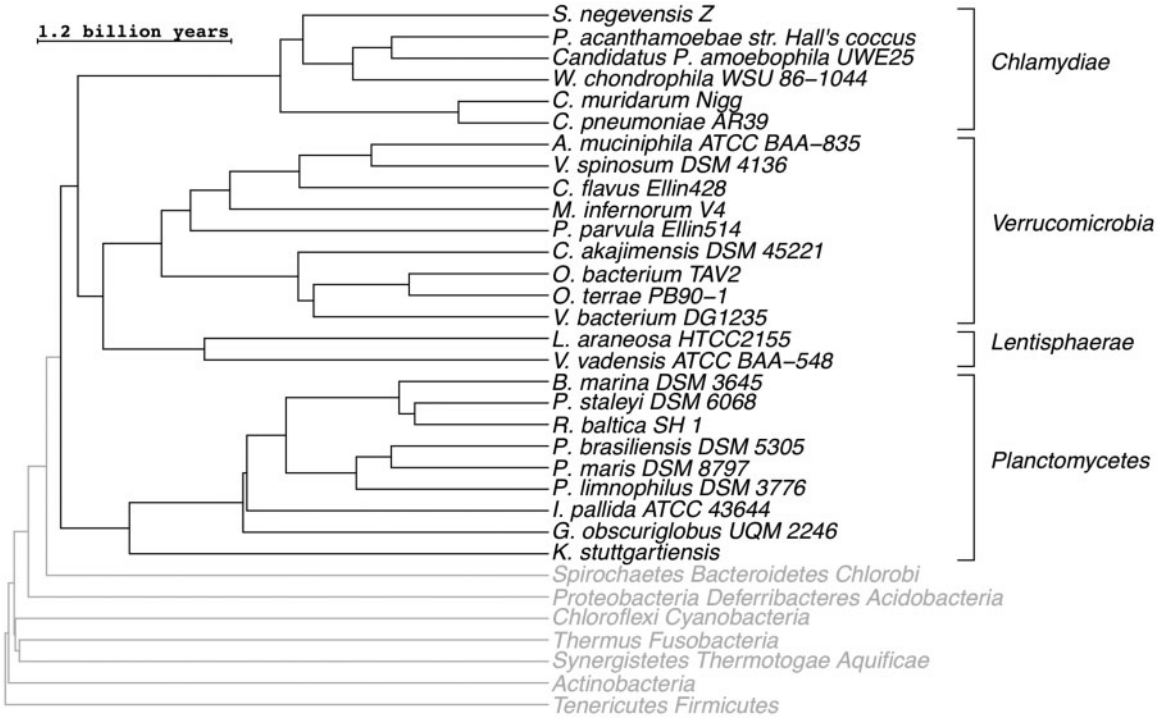

Fig. 1.—Evolutionary relationships between PVC organisms included in the analysis. Species tree topology was recovered for the entire set of 99 bacterial species from various bacterial phyla, as a consensus tree averaging over gene trees of 44 phylogenetic markers. Divergence times were estimated using a concatenated alignment of all 44 phylogenetic markers. Non-PVC clades were collapsed for clarity and corresponding lineages are shown in gray. Names of PVC phyla are shown on the right. The full species tree is presented in supplementary figure S1, Supplementary Material online.

phylogenetic relationships between bacterial species. Patchiness in recombination and frequent HGT events are often apparent when clear phylogenetic separation or coherence of species can be established using one set of markers, whereas other sets support alternate groupings (Hanage et al. 2005; Fraser et al. 2007).

Evolution of the Compartmentalized Cell Plan in the PVC Super-Phylum

One of the questions motivating the study of the evolutionary history of the PVC super-phylum is the origin of PVC-specific intracellular membranes. The lineage or lineages where the genetic determinants of PVC-specific cellular structures were acquired is unknown. To address this question, we performed ancestral character reconstruction using the model described in (Pagel 1999). We inferred the presence/absence of intracellular membranes in ancestral species corresponding to the nodes of the species tree (supplementary fig. S10, Supplementary Material online) using data on the presence/absence of intracellular membranes obtained from published studies (Fuerst 2005; Lee et al. 2009, Fuerst JA, personal communication) (supplementary table S1, Supplementary Material online). According to our inference, the compartmentalized cell plan is likely to have initially emerged on the PVC phylogenetic lineage; therefore, the genes responsible for the development of cellular compartmentalization should also have been first acquired/emerged on this phylogenetic lineage within the super-phylum. It is also possible that the compartmentalized cell plan was reinvented on some internal lineages of *Verrucomicrobia*- and lentisphaerae-specific clades, but additional microscopy experiments are needed to elucidate if and when that happened.

Gene Family Dynamics in the PVC Super-Phylum

We employed OrthoMCL clustering to derive putative gene families. We identified 17,608 homologous families containing at least three sequences, 9,444 containing two members, and 63,679 singletons. To assess gene family dynamics, we reconciled every gene tree with the species tree using the parsimony-based method reported in (David and Alm 2010). The algorithm uses HGT, gene duplication, and gene loss events to explain discrepancies between gene and species phylogeny. Every event is associated with a penalty, from which the algorithm then infers reconciliation of minimal cost. The use of varying event penalties within a parsimony-based gene-tree species-tree reconciliation procedure allows the recovery of every possible reconciliation as the most parsimonious one. Therefore, careful justification of event penalties is required. We conducted a search for HGT and duplication cost values which lead to minimum average genome flux (the difference between the ancestor and descendant genome sizes over every lineage of the species tree). We used genome flux as an external criterion to evaluate the accuracy of every inferred evolutionary scenario of genome-level changes. The main assumption of this step was that the genome sizes of the ancestor and descendant are correlated. This method was also employed in the original article reporting use of the AnGST tool (David and Alm 2010). Although genome sequences are expected to harbor segregating variations along with sites invariant across the entire population, we treated all genomes as if they represent a genome of the entire species. This was due to the lack of population data for the majority of species under consideration, and was thus a necessary assumption of this analysis.

Gene-tree species-tree reconciliation allows explicit inference of the evolutionary history of every gene family in the data set, in terms of gene family origins, gene duplication, loss, and transfer events. It also allows evaluation of genome size for every ancestral genome. We additionally conducted Bayesian statistical modeling of evolutionary rates (see Materials and Methods section for details) and identified branches with accelerated and decelerated rates of gene loss, duplication, birth, and acquisition via HGT. The summarized results of this analysis for the PVC clade are depicted in figure 2; data from the full data set are illustrated in supplementary figure S11, Supplementary Material online. Here, we identified potential HGT events based on gene-tree species-tree reconciliation. However, this assertion should be viewed as "the best guess" for the origin of the sequence in the genome considering the limited number of extant lineages included in the data set, as well as the mosaic nature of the majority of studied genomes (McInerney et al. 2011).

Our results suggest that the common ancestor of all PVC organisms had a genome containing 3,106 genes, of which 786 were members of multigene families. Three hundred fifty-nine gene families emerged on the lineage leading to the common ancestor of all PVC organisms. Only a small proportion (7) of these 359 families encodes membrane coat-like proteins; therefore, other gene families acquired on this lineage could contain previously uncharacterized determinants of the PVC-specific cell plan. Two hundred twenty-nine gene families were acquired via HGT on the lineage leading to the last common ancestor of all PVC species.

After the *Planctomycetes* split from the rest of the superphylum, lineages leading to *Rhodopirellula baltica*, *Blastopirellula marina*, *Planctomyces maris*, and the common ancestor of *Planctomyces* and *Pirellula* species underwent birth of gene families. We also observed elevated gene birth rate, and expansion of existing families through gene duplication, on the lineage leading to *G. obscuriglobus*, with large numbers of genes in multigene families (2,840 genes out of 7,989). This suggests that evolution on the *G. obscuriglobus* lineage has been affected by changes after gene duplication, although we did not test this hypothesis; additional factors affecting duplicate retention include dosage balance, when the gene is retained due to required stoichiometric balance with other proteins in the system, and selection on abundance

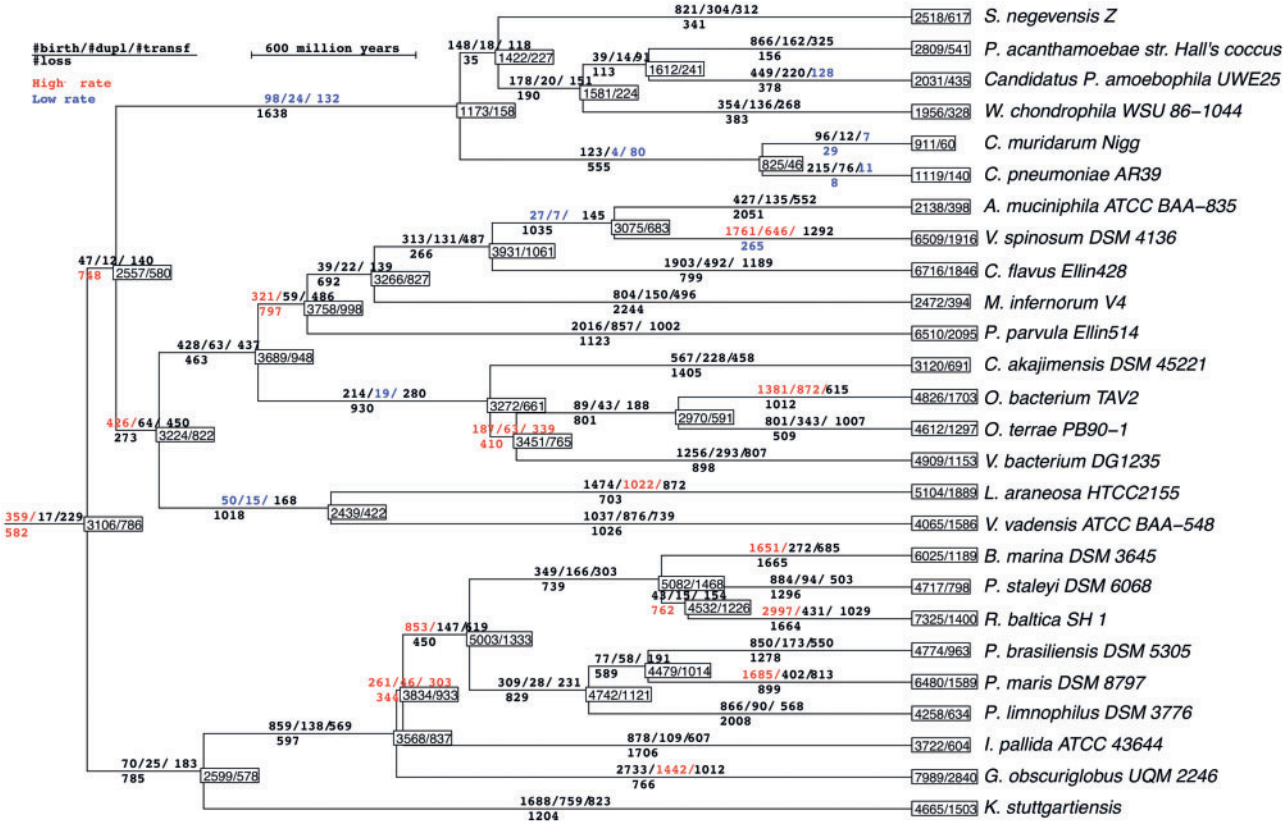

Fig. 2.—Genome content evolution in PVC super-phylum. Events of genome content evolution were mapped onto lineages of the species tree; only the PVC clade is shown here. Numbers at every node, either ancestral or extant, represent genome size and number of genes in multigene families (for instance the *Isosphaera pallida* genome contains 3,722 genes, out of which 604 genes are predicted to be members of multigene families). Numbers above and below every lineage represent the number of birth/duplication/transfer and loss events, respectively, predicted to occur on the branch. Numbers shown in red or blue correspond to accelerated or decelerated rates of events on the branch, classified as described in Materials and Methods section (for instance on the lineage leading to *Verrucomicrobium spinosum* 1,761 births, 646 duplications, 1,292 transfers, and 265 loss events occurred. This observed event count implies elevated gene birth and duplication rates on this lineage and low gene loss rate).

of gene product of nondivergent duplicates (Innan and Kondrashov 2010; Dittmar and Liberles 2011). The ancestor of *Verrucomicrobia*, *Chlamydiae*, and *Lentisphaerae* possessed a relatively large genome containing 2,557 genes. This ancestral gene set was shaped primarily by genome shrinkage on the lineage leading to the ancestor of all the *Chlamydiae*, and within the chlamydial clade. We did not detect elevated rates of gene loss in the chlamydial clade, but rather found evidence for low rates of gene family acquisition through various mechanisms. Generally, low rates of gene birth, duplication, and transfer were detected on several lineages of the chlamydial clade, especially on the lineage leading to the common ancestor of obligate intracellular pathogens of the genera *Chlamydia* and *Chlamydophila*. Gene gain via birth of new gene families prevailed on the lineage leading to the common ancestor of *Lentisphaerae* and *Verrucomicrobia*. This was followed by deceleration of gene birth and duplication rates on the lineage leading to the common ancestor of *Lentisphaerae*. Subsequently, an accelerated gene duplication process occurring on the *Lentisphaera araneosa* lineage resulted in the high number of genes in multigene families in the *L. araneosa* genome, again pointing to the role of protein evolution after gene duplication. Gene gain contributed to genome content changes on several lineages of phylum *Verrucomicrobia* (*Verrucomicrobium spinosum*, *Opitutaceae bacterium*, and several ancestral lineages). Low rates of gene birth and duplication were detected on the lineage leading to the common ancestor of *V. spinosum* and *A. muciniphila*, which may have played a role in the emergence of the relatively small genome of *A. muciniphila* found in close association with the eukaryotic host.

We also detected patterns consistent with a high degree of genome plasticity and possibly emergence of new phenotypes on several phylogenetic lineages (with elevated rates of processes bringing new genes into the genome [gene birth and HGT] and gene loss for the same lineage). We observed this pattern for a number of ancestral lineages within the super-phylum including the following: lineages leading to

last common ancestors of all PVC organisms; *O. bacterium*, *Opitutus terrae*, and *Verrucomicrobiae bacterium; A. muciniphila*, *V. spinosum*, *Chthoniobacter flavus*, *Methylacidiphilum infernorum*, and *Pedosphaera parvula;* species belonging to order *Planctomycetales.* For two of those lineages, elevated HGT rates are also consistent with a highly chimerical nature of the emerging genome.

This report summarizes for the first time genome-level processes that appear to have occurred on various lineages of the PVC super-phylum. Here, we determined that genome expansion has driven evolution within *Planctomycetes*, genome contraction has been a main theme of genome-level changes within *Chlamydiae*, and the balance of these two processes has led to genomes of stable size in *Verrucomicrobia* and *Lentisphaerae.* We observed acceleration of gene birth and gene loss rates on the PVC evolutionary lineage, which suggests that genome content changed significantly in terms of present/absent genes, to produce the common ancestor of the super-phylum. This indicates that a number of new phenotypes might have emerged on this lineage.

Horizontal Gene Transfer among PVC Organisms and Members of Other Bacterial Groups

Lateral transfer is recognized as a major source of diversity in bacterial genomes (Jain et al. 2003). However, an accepted model of HGT which would allow prediction of the frequency of transfer events between a given pair of species does not currently exist. This frequency is known to depend on differences in GC content and genome size, as well as ecological factors such as carbon source utilization and oxygen tolerance (Jain et al. 2003; Smillie et al. 2011). Other factors that intuitively should affect frequency of HGT between organisms include differences in codon usage, di-, and tri-nucleotide composition, and divergence between donor and recipient organisms, including divergence of regulatory motifs and biomolecular interaction interfaces. The exact mode of physical interaction in an ecosystem (whether organisms are physically attached to each other and how they might interact, versus rare or nonexistent physical contact) must also be critical to HGT. However, the modes of interaction between bacterial cells in the environment are hard to assess. To delineate the extent to which HGT contributes to PVC genome dynamics, we traced the number of transfer events between different organisms for every gene family within the data set (table 1, supplementary table S3 and fig. S12, Supplementary Material online). We then determined which lineages were overrepresented among donors of transferred genes for every lineage of the species tree. We used a simple binomial test with the probability of an organism (extant or ancestral) acting as an HGT donor being proportional to its genome size, but assuming that all transfer events were independent and all genomes had equal opportunity to contribute. For simplicity, we did not account for all the other factors affecting HGT frequency. Deviation from the random process should therefore be viewed as a sign that one of the factors listed above has affected HGT between the species. It appeared that HGT has preferentially occurred between organisms of the same phylum, which is attributable to the existence of fewer divergent regulatory mechanisms within a group of related species and general similarity in genome structure. It is also possible that lifestyle properties are not radically different between closely related bacterial taxa and therefore co-occurrence in the environment might also contribute. Three main HGT highways, all occurring within *Planctomycetes*, were detected within the PVC clade: a large number of sequences were transferred from *P. staleyi* to *B. marina* (149 events), and from *P. maris* and *Planctomyces brasiliensis* to *P. limnophilus* (120 and 146 events, respectively). The species in the list are referred to as "donor organisms" for simplicity; however, they

Table 1
Organisms (Extant or Ancestral) Frequently Acting as Donors in Lateral Transfer Events

Recipient	Donor (No. of Transfer Events); Only Organisms Frequently Acting as Donors Are Shown ($P < 1e{-}8$)
Kuenenia stuttgartiensis	Deferribacteres (18); *Desulfovibrio vulgaris* Miyazaki F (25); *Geobacter lovleyi* SZ (39); *Synergistetes* (15); *Hydrogenobacter/Persephonella/Sulfurihydrogenibium* (15)
Gemmata/Isosphaera/Pirellulaceae/Planctomyces	Candidatus *Solibacter usitatus* Ellin6076 (45)
Gemmata obscuriglobus UQM 2246	Candidatus *S. usitatus* Ellin6076 (49)
Isosphaera pallida ATCC 43644	*Chloroflexus aggregans* DSM 9485 (22)
Victivallis vadensis ATCC BAA-548	*Spirochaeta* sp. Buddy (21); *Treponema azotonutricium* ZAS-9 (22); *D. vulgaris* Miyazaki F (25)
Methylacidiphilum infernorum V4	a-Proteobacteria (22)
Pedosphaera parvula Ellin514	*T. saanensis* SP1PR4 (34); Candidatus *S. usitatus* Ellin6076 (78)
Opitutaceae/Opitutus/Verrucomicrobiae	Candidatus *S. usitatus* Ellin6076 (29)
Verrucomicrobiae bacterium DG1235	Candidatus *S. usitatus* Ellin6076 (38)
Opitutus terrae PB90-1	Candidatus *S. usitatus* Ellin6076 (81); *D. vulgaris* Miyazaki F (24)
Chthoniobacter flavus Ellin428	*Sorangium cellulosum* So ce 56 (60); Candidatus *S. usitatus* Ellin6076 (63)
Akkermansia muciniphila ATCC BAA-835	*Bacteroides fragilis* NCTC 9343 (47); *D. vulgaris* Miyazaki F (17)

reflect merely the best guesses from currently available genomic sequences. Interpretation of the data assumes that lifestyle properties are not radically different between closely related bacterial taxa if the true donor is not included in the data set.

There was also evidence for HGT events between phyla. Identification of HGT donors external to the PVC superphylum could provide insights into the evolution of the recipient's ecology and cell biology, which is especially relevant for these poorly characterized organisms. Although the ecological niche of *V. spinosum* is unknown, it has been proposed that this organism is capable of host associations (Sait et al. 2011). According to our analysis, a number of genes in the *V. spinosum* genome were derived from HGT events from organisms outside the super-phylum, including Candidatus *Solibacter usitatus* Ellin6076 (*P* value = 9e−07, 42 genes), *Desulfovibrio vulgaris* Miyazaki F (*P* value = 5e−06, 22 genes, including a number of genes encoding type III secretion), and *Sorangium cellulosum* So ce 56 (*P* value = 8.972e−04, 38 genes). Although these donor organisms are typically considered free-living, *Desulfovibrio* species are also known to be host associated, and in some cases pathogenic (Goldstein et al. 2003). This suggests that *V. spinosum* might also exhibit a host-associated lifestyle; however, the mode of this interaction is unknown. In contrast, *G. obscuriglobus*, a free-living aquatic organism (Franzmann and Skerman 1984), also acquired (according to our computational analysis) a large number of genes via HGT from Candidatus *S. usitatus* Ellin6076 (*P* value < 1e−08, 49 genes), *Sor. cellulosum* So ce 56 (*P* value = 2e−07, 42 genes), and the ancestor of all alphaproteobacteria (*P* value = 4.23e−05, 17 genes), but only 12 genes from *D. vulgaris* Miyazaki F (*P* value = 8.5e−03).

In contrast to the quite high HGT activity for these (primarily) free-living lineages, we detected a very small number of HGT events involving chlamydial species, which highlights the consequences of the host-dependent lifestyle of *Chlamydiae* and the distinct evolutionary forces acting on their genomes. However, five HGT events were detected between the ancestor of all *Chlamydiae* and the ancestor of spirochaetes (*P* value = 9.21e−06) and an additional five genes were acquired from the common ancestor of cyanobacteria (*P* value = 5.56e−06).

Here, we characterized the contribution of HGT to PVC super-phylum evolution in a quantitative manner, and determined that, according to our analysis, many genes were acquired by extant and ancestral PVC organisms from *Acidobacteria* (in particular from Candidatus *S. usitatus*), delta *Proteobacteria* (in particular from *Sor. cellulosum* and *D. vulgaris*) and, in the case of *A. muciniphila*, from *Bacteriodetes* (in particular from *Bacteroides fragilis* NCTC 9343, described later). Large numbers of observed transfer events between PVC species and other bacterial organisms presumably indicate close ecological interactions between bacteria in the environment.

Gene Birth, Duplication, Transfer, and Loss in Different Biological Functions

Major evolutionary events are observed at the level of the individual gene/protein but act in the context of broader cellular biology. We used KEGG (Kanehisa et al. 2010) metabolic pathways to classify gene families and systematically identify pathways affected by gene birth, duplication, transfer, and loss processes. We elucidated overrepresented/underrepresented pathways among gene families affected by these events. We found that gene birth predominantly affected genes of unknown function (according to KEGG) and other pathways were significantly depleted; cluster 1 mostly contained lineages leading to extant species (supplementary fig. S13*A*, Supplementary Material online). However, within cluster 1, there were some outliers, for instance cell motility genes were overrepresented among genes born on the lineage leading to the common ancestor of *Chlamydia* and *Chlamydophila*. Several ancestral lineages of different phyla form cluster 2, where some pathways were affected by gene birth to different extents. For example, pathways for metabolism of terpenoids, polyketides, and other secondary metabolites were enriched among genes born on lineages leading to the common ancestor of Lentisphaerae species and the *Chthoniobacter/Methylacidiphilum/Akkermansia/Verrucomicrobium* phylogenetic lineage. Proteins involved in folding, sorting, and degradation were affected by gene birth on *Chlamydiae*, *Lentisphaerae/Verrucomicrobia/Chlamydiae*, and *P. maris/P. brasiliensis* lineages. Cluster 3 includes mostly ancestral lineages where genes involved in transcription and genes of unknown function were considerably affected by gene birth.

In cases of HGT, we were able to distinguish three main clusters of phylogenetic lineages (supplementary fig. S13*B*, Supplementary Material online). In cluster 1, HGT did not affect genes of unknown function but instead influenced the majority of other pathways. In cluster 2 and 3, genes of unknown function, lipid metabolism, and xenobiotic biodegradation and metabolism were strongly affected by HGT. In cluster 2, genes involved in cell motility and signal transduction were affected, whereas in the majority of lineages from cluster 3, membrane transport, glycan biosynthesis and metabolism, metabolism of other amino acids, and biosynthesis of other secondary metabolite pathways were influenced by HGT.

We identified three major clusters of evolutionary lineages based on how genes in various functional categories changed via gene duplication on these lineages (supplementary fig. S13*C*, Supplementary Material online). On lineages from cluster 3, gene duplication mainly affected genes of unknown function. Other functional categories were significantly underrepresented, with the exception of signal transduction and cell motility pathways on lineages leading to several extant or ancestral lineages within *Verrucomicrobia*- and *Planctomycetes*-specific clades. In cluster 2, genes of unknown function were

significantly overrepresented among duplicated genes. However, the remaining functional categories were not depleted but rather affected at average rate. In cluster 1, genes of unknown function were not affected by gene duplications, whereas a number of general metabolic and genetic information processing functions were. With regard to how gene loss influences various cellular functions on different evolutionary lineages within the PVC super-phylum, we distinguished three main groups of phylogenetic lineages (supplementary fig. S13*D*, Supplementary Material online). Cluster 2 mostly included extant and ancestral lineages of the *Chlamydiae*-specific clade; the most affected cellular functions in this cluster were amino acid metabolism and energy metabolism, which supports our current understanding of genome evolution in phylum *Chlamydiae* (Kalman et al. 1999).

Enrichment of unclassified genes (unassigned to specific pathways) among those influenced by particular evolutionary events, as well as the generally large number of unclassified gene families within PVC genomes, might indicate a large number of uncharacterized pathways evolved within PVC species.

HGT Example: Carbohydrate Utilization in *A. muciniphila*

Organisms inhabiting the same environment tend to have a large number of shared genes; this process is driven by adaptation to the same ecological conditions of the habitat as well as greater opportunity for HGT due to close physical contact of cells. Consistent with this, a large number of transfer events were detected between *A. muciniphila* and *B. fragilis* (organisms associated with the human digestive tract), 47 genes, P value $< 1e{-}08$. Both of these organisms are capable of mucin degradation (Roberton and Stanley 1982; Derrien et al. 2004) and *A. muciniphila* is the only confirmed host-associated *Verrucomicrobia* organism included in our analysis; therefore, all determinants of this new lifestyle should be acquired on the *A. muciniphila* lineage. Horizontal acquisition of carbohydrate-binding peptidases (proteins implicated in mucin degradation) by *A. muciniphila* (and other PVC organisms) from *Bacteroidetes* species was recently reported (Nakjang et al. 2012). However, the evolutionary origin of pathways related to carbohydrate metabolism in *A. muciniphila* has not been described. To address this question and explore one example where evolutionary process affects molecular function but is acting on the level of particular genes, we performed a comprehensive phylogenomic analysis of *A. muciniphila* genes predicted to encode carbohydrate-active enzymes by CAZY classification (http://www.cazy.org, last accessed December 13, 2012) (Cantarel et al. 2009). There are four carbohydrate esterases (three families); 59 glycoside hydrolases (26 families); and 44 glycoside transferases (14 families) encoded in the genome of *A. muciniphila*. Forty-six out of 59 putative glycoside hydrolyses possess a predicted signal peptide (LipoP or SignalP prediction) and therefore might be potentially involved in mucin degradation outside the bacterial cell (supplementary table S4, Supplementary Material online).

We inferred the evolutionary origin of every glycoside hydrolase and transferase gene using a phylogenomics framework. Phylum-level taxonomic classification of organisms from the closest sister clade was used to identify potential origin for every putative glycoside hydrolase or glycoside transferase in the *A. muciniphila* genome (fig. 3). Three groups of glycoside hydrolase families were observed when classified based on the taxonomic affiliation of phylogenetically closely related sequences. Enzymes of the first group were vertically transmitted to *A. muciniphila* and group together with other sequences from verrucomicrobial genomes; they belong to families 98, 97, 77, 3, and 18. Proteins from the second group were closely related to genes from *Bacteroidetes* species, glycoside hydrolase families 110, 27, 13, 57, 31, 43, 123, 105, 109, 29, 2, and 20. The third group included glycoside hydrolase families represented by 2 to 4 genes in *A. muciniphila*. One copy was affiliated with a protein from *Bacteroidetes* species, and the others either vertically transmitted to *A. muciniphila*, or shared an origin with proteins in *Acidobacteria* genomes. Glycoside hydrolase families of the fourth group were represented by a few (1–3) proteins in the genome and according to our classification, closely related to proteins from organisms of different taxa or their ancestral species (*Lentisphaerae*, *Actinobacteria*, *Acidobacteria*, or *Planctomycetes*). Families of glycoside hydrolases known to be involved in mucin degradation (families 27, 29, 2, 20, 95, 35, 36, and 33) (Turroni et al. 2010) were present in group number 2 and 3. This suggests a dual evolutionary origin of mucin degradation pathways in *A. muciniphila*, where genes in the pathway were obtained from two major donors, which also gave rise to sequences in *Acidobacteria* and *Bacteroidetes*.

A different trend was observed in the case of glycoside transferases, with the main group of glycoside transferase families containing genes vertically inherited by *A. muciniphila*. The second cluster contained glycoside transferases of varying taxonomic affiliation (distantly related *Chloroflexi*, *Synergistetes*, *Proteobacteria*, *Bacteroidetes*, or more closely related *Planctomycetes*). This pattern suggests vertical transition for the majority of genes involved in carbohydrate biosynthesis. The difference between glycoside transferases and glycoside hydrolases suggests that glycan utilization is a major mechanism of microbial survival within the human digestive system and highlights the importance of extracellular glycoside degradation for *A. muciniphila* fitness.

Gene Birth Example: Proteins Containing the DUF1501 and PSCyt Domains in *Planctomycetes, Verrucomicrobia*, and *Lentisphaerae*

The discovery of cellular compartmentalization within first *Planctomycetes*, and subsequently *Poribacteria* and

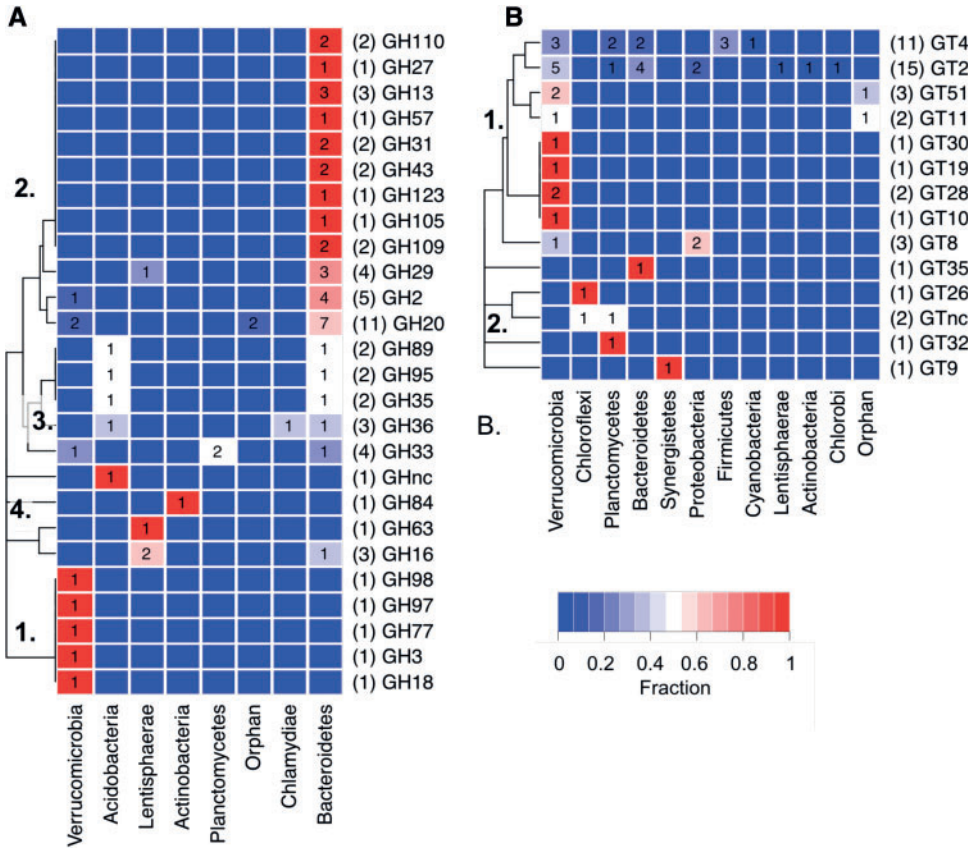

Fig. 3.—Taxonomic assignment of sequences closely related to carbohydrate metabolism genes from *Akkermansia muciniphila*. (*A*) Glycoside hydrolases; (*B*) Glycoside transferases. Heatmap chart was generated using ratios of genes with a certain taxonomic assignment to total number of genes assigned to a glycoside hydrolase or transferase family. Dendrograms on the sides of the heatmaps were generated using a hierarchical clustering algorithm. Clusters discussed in the main text are marked. Color key is present in the bottom-right corner.

Verrucomicrobia, established the super-phylum as a model group for studying the evolution of biological complexity within bacteria (Fieseler et al. 2004; Fuerst 2005; Lee et al. 2009). The identification of proteins necessary for the transition from a simple bacterial cell plan to a more complex compartmentalized cellular organization is of great importance to evolutionary biology. Although some genes have been implicated in their role in cellular compartmentalization within the super-phylum (Santarella-Mellwig et al. 2010), genetic determinants of this process are yet to be determined. Within this study, we identified several gene families which were detected primarily within the PVC super-phylum and asserted to have emerged on this evolutionary lineage (supplementary table S5, Supplementary Material online). These genes might be involved in compartment formation within PVC species, although they could also be present in the genomes due to species relatedness. Several identified gene families contained proteins previously reported to be PVC-specific membrane coat-like proteins (Santarella-Mellwig et al. 2010). The others contained protein domains of unknown function. One such domain is DUF1501 (PF07394). DUF1501-containing proteins were present in *Planctomycetes* (*I. pallida*, *G. obscuriglobus*, *Planctomyces*, and *Pirellula* species), *Verrucomicrobia* (*Par. parvula*, *V. spinosum*, *C. flavus*, and *Coraliomargarita akajimensis*), and *Lentisphaerae* (*L. araneosa*) genomes, as well as a few species of *Bacteroidetes*, *Acidobacteria*, and *Proteobacteria* (supplementary fig. S14, Supplementary Material online). All the organisms in this list either have been shown to contain intracellular membranes or have not been analyzed in terms of their cellular structure. DUF1501-containing proteins are absent from the basal *Planctomycetes* species, *Kuenenia stuttgartiensis*, which is known to possess intracellular membranes. This suggests that DUF1501 is not related to the primary emergence of intracellular membranes within PVC but might be involved in processes that are associated with the intracellular membranes in more recently derived *Planctomycetes* and *Verruconicrobia*. It might also indicate a distinct genetic basis for compartmentalization in "anammox" species compared with organisms of the class *Planctomycetia*. The large number of DUF1501-containing proteins indicated that they are involved in a variety of related cellular processes and are

important for the organisms; therefore we carried out further characterization of DUF1501-containing proteins.

To identify proteins or protein domains functionally related to DUF1501, we conducted analysis of the genomic neighborhood of all DUF1501-containing proteins. This analysis revealed the existence of genomic clusters of varying structure and domain composition (fig. 4 and supplementary fig. S14, Supplementary Material online). The proteins most strongly associated with DUF1501 featured a cytochrome c motif containing domains PSCyt1 and PSCyt2 (PF07635 and PF07583) and domain of unknown function PSD1 (PF07587), previously identified to be specific for *R. baltica* (Studholme et al. 2004). Figure 4 shows the domain composition of several genomic clusters containing DUF1501 proteins found in the genome of *V. spinosum*, present in a variety of other PVC species and showing clear grouping within the phylogenetic tree of DUF1501-containing proteins (supplementary fig. S14, Supplementary Material online). Some DUF1501-containing proteins were not associated with proteins bearing PSCyt or PSD1 domains. Another genomic locus architecture included a DUF1501-containing protein (sometimes also carrying a twin-arginine motif or signal peptide) and a protein with PSCyt1/PSCyt2/PSD1 architecture. The second protein in the genomic cluster sometimes contained additional Laminin_G_3 or F5_F8_type_C domains classically involved in carbohydrate binding (Sharon and Lis 1972). The most complex gene clusters involved four genes with the following predicted domain architecture: 1) DUF1501, sometimes with twin-arginine signal peptide; 2) protein with weak support for one or several PPC domains normally found in secreted bacterial peptidases (Yeats et al. 2003), signal peptide, and conserved regions without characterized signatures; 3) PSCyt1/Big_2/PSCyt2/PSD1 protein; and 4) PSCyt1/WD40 protein. Domains Big_2 and WD40 are known to be involved in protein–protein interactions (Kelly et al. 1999; Xu and Min 2011) and probably are responsible for protein complex assembly or substrate recognition. The correlation coefficient between the presence/absence of intracellular membrane and the phylogenetic distribution of genomic loci of this structure was 0.852, indicating a strong association of this genetic module with intracellular membranes. Several organisms of undetermined cellular structure also encode the genomic clusters, suggesting that they might also possess internal membranes.

The large number of genomic clusters, diversity of their domain composition, and high degree of evolutionary conservation of DUF1501-containing genomic loci suggests that DUF1501, PSCyt1/2, and PSD1 domain containing proteins are involved in a variety of related cellular processes and are important for the organisms. From an evolutionary standpoint, the topology of the large-scale phylogenetic tree for DUF1501 containing proteins shows clusters of DUF1501-bearing genes from the loci of similar domain architecture (supplementary fig. S14, Supplementary Material online). DUF1501-containing proteins are preferentially present within *Planctomycetes*, *Verrucomicrobia*, and *Lentisphaerae*. This suggests that this protein domain 1) has emerged on the lineage leading to the common ancestor of PVC organisms, 2) duplicated a number of times and acquired a functional relationship with a variety of proteins on the PVC evolutionary lineage, and 3) has been subsequently transmitted vertically through the PVC clade of the species tree with a number of loss events occurring within the clade. This implies that the majority of the observed domain arrangements for DUF1501-containing genomic loci have emerged on the PVC evolutionary lineage in a time of ~100 Myr, constituting a rapid phase of functional innovation within DUF1501-associated genetic modules. This is consistent with a currently accepted model of protein evolution where domain rearrangements play a major role in evolutionary adaptations (Bornberg-Bauer et al. 2010). Some domain arrangements appear to be reinvented within a number of genomic loci (supplementary fig. S14, Supplementary Material online), which is in agreement with the current view of convergent evolution of domain architectures being rather common (Forslund et al. 2008).

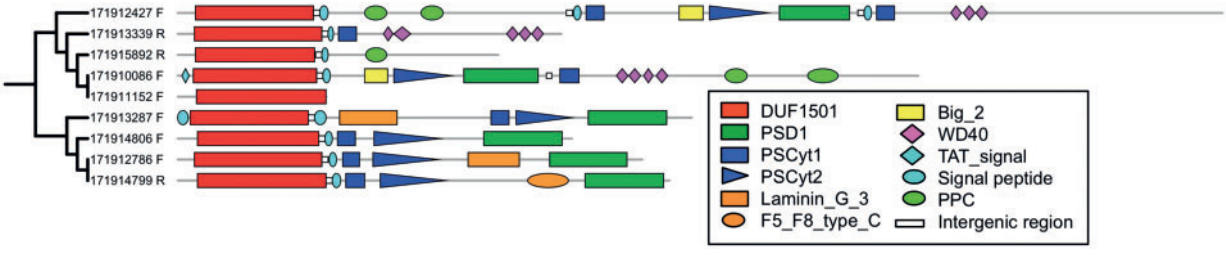

Fig. 4.—Domain structure of several different types of genomic loci including proteins that contain DUF1501, PSCyt1/2, and PSD1 domains in the *Verrucomicrobium spinosum* genome. Evolutionary relationships between several DUF1501 domain-containing proteins from the *V. spinosum* genome, representing genomic loci of different architecture conserved across several species, are shown. GI numbers of corresponding *V. spinosum* proteins are used as tip labels. Forward and reverse orientation in the genome are denoted by F or R sign. Domain architecture is shown for the genomic neighborhood of every DUF1501 domain-containing protein. Codes for the different domains are shown in the figure inset. Large-scale phylogeny for all DUF1501-containing proteins identified in all completely sequenced bacterial genomes, and complete or draft PVC genomes, is shown in supplementary figure S13, Supplementary Material online.

Association of this system with intracellular membrane compartmentalization (within the class *Planctomycetia* and other PVC species) is additionally supported by the presence of canonically membrane-associated putative cytochromes and signal peptides within the described units. Additional intracellular membranes in this case might constitute alternative destinations for canonically outer membrane-targeted proteins. The presence of a number of carbohydrate- and protein-binding units within PSCyt domain-carrying genes suggests that those genes encode outer membrane-associated (or according to our hypothesis, intracellular membrane-associated) cytochromes transferring electrons to specific acceptors (possibly proteins and sugars). At the same time, the presence of a twin-arginine signal peptide within DUF1501-bearing proteins points to an enzymatic function, which requires cofactors acquired in the cytoplasm of the cell (Lee et al. 2006) and carried out either in the periplasm (or possibly within additional intracellular compartments) or outside the cell. However, the function itself remains unclear.

Overall, the described ensemble of genetic modules constitutes an interesting example of a novel molecular system within a number of PVC genomes. The high level of evolutionary conservation and the number of times the clusters have been duplicated in the genomes suggests an important and lineage-specific role of encoded proteins in the bacterial cell. It also implicates involvement of DUF1501 and PSCyt/PSD-containing proteins in a large number of cellular processes. All these make this system an attractive target for future functional studies. Recent breakthroughs in developing genetic tools for *Verrucomicrobia* and *Planctomycetes* species (Domman et al. 2011; Jogler et al. 2011) provide opportunities to test hypotheses presented here.

Conclusions

Here, we uncovered several features of PVC genome evolution. First, organisms in the super-phylum evolved primarily via genome shrinkage within the chlamydial clade, genome expansion within the Planctomycete clade, and a balance of these two processes on the lineages within *Verrucomicrobia* and *Lentisphaerae*. Second, acquisition of novel genes (gene birth) has been a consistent characteristic of PVC genome evolution, but the rate of gene birth varies among lineages of the clade and the rate of gene acquisition has increased on a variety of recent *Planctomycetes* lineages and on the ancestral PVC lineage. Third, large numbers of genes were acquired by extant and ancestral PVC organisms from *Acidobacteria*, delta *Proteobacteria* and, in the case of *A. muciniphila*, from *Bacteroidetes*.

We also described an example of a complex molecular system that evolved within the PVC super-phylum. Considering the predicted evolutionary origin, phylogenetic distribution, and revealed evolutionary patterns, the system is likely to be involved in multiple processes and be uniquely important for PVC bacteria bearing those genes. The phylogenetic distribution of the proteins seems to correlate nondeterministically with the presence/absence of the additional intracellular membrane in *Planctomycetes* and *Verrucomicrobia* (correlation coefficient of 0.852 for the DUF1501-containing protein family from the most complex four-gene genomic cluster). Further study of these novel proteins might provide insights into the emergence of complex cellular structures within bacteria. Ultimately, the combination of high-throughput analysis and detailed analysis of individual protein families constitutes a useful approach for understanding the evolution of bacterial genomes.

Supplementary Material

Supplementary figures S1–S14 and tables S1–S5 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

The authors thank C. Alex Buerkle, Miklós Csűrös, Donald Jarvis, Oleg Moskvin, Heather Rothfuss, Corrine Seebart, and Ekaterina Yarunova for helpful discussions. This work was supported by the National Institutes of Health (P20 RR016474 to O.K.K. and S.J.K.) and National Science Foundation (DBI-0743374 to D.A.L. and MCB-0920667 to N.L.W.). N.L.W. and O.K.K. were also partially supported by NSF EPS-0447681. The content of this paper is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Literature Cited

Abascal F, Zardoya R, Posada D. 2005. ProtTest: selection of best-fit models of protein evolution. Bioinformatics 21:2104–2105.

Åkerborg O, Sennblad B, Arvestad L, Lagergren J. 2009. Simultaneous Bayesian gene tree reconstruction and reconciliation analysis. Proc Natl Acad Sci U S A. 106:5714–5719.

Altschul S, Gish W, Miller W, Myers E, Lipman D. 1990. Basic local alignment search tool. J Mol Biol. 215:403–410.

Anisimova M, Liberles D. 2007. The quest for natural selection in the age of comparative genomics. Heredity 99:567–579.

Balows A, Trüper HG, Dworkin M, Harder W, Schleifer KH. 1992. The prokaryotes: a handbook on the biology of bacteria: ecophysiology, isolation, identification, applications. New York: Springer.

Benson D, Karsch-Mizrachi I, Lipman D, Ostell J, Wheeler D. 2005. GenBank. Nucleic Acids Res. 33:D34–D38.

Berglund-Sonnhammer A, Steffansson P, Betts M, Liberles D. 2006. Optimal gene trees from sequences and species trees using a soft interpretation of parsimony. J Mol Evol. 63:240–250.

Bornberg-Bauer E, Huylmans A-K, Sikosek T. 2010. How do new proteins arise? Curr Opin Struct Biol. 20:390–396.

Britten R. 2002. Divergence between samples of chimpanzee and human DNA sequences is 5%, counting indels. Proc Natl Acad Sci U S A. 99:13633–13635.

Cantarel B, et al. 2009. The Carbohydrate-Active EnZymes database (CAZy): an expert resource for glycogenomics. Nucleic Acids Res. 37:D233–D238.

Ciccarelli F, et al. 2006. Toward automatic reconstruction of a highly resolved tree of life. Science 311:1283–1287.
Cohen O, Gophna U, Pupko T. 2010. The complexity hypothesis revisited: connectivity rather than function constitutes a barrier to horizontal gene transfer. Mol Biol Evol. 28:1481–1489.
Csűös M. 2010. Count: evolutionary analysis of phylogenetic profiles with parsimony and likelihood. Bioinformatics 26:1910–1912.
David L, Alm E. 2010. Rapid evolutionary innovation during an Archaean genetic expansion. Nature 469:93–96.
Davids W, Gamieldien J, Liberles D, Hide W. 2002. Positive selection scanning reveals decoupling of enzymatic activities of carbamoyl phosphate synthetase in *Helicobacter pylori*. J Mol Evol. 54:458–464.
Derrien M, Baarlen PV, Müller M. 2011. Modulation of mucosal immune response, tolerance, and proliferation in mice colonized by the mucin-degrader *Akkermansia muciniphila*. Front Microbiol. 2:166.
Derrien M, Vaughan EE, Plugge CM, de Vos WM. 2004. *Akkermansia muciniphila* gen. nov., sp. nov., a human intestinal mucin-degrading bacterium. Int J Syst Evol Microbiol. 54:1469–1476.
Dittmar K, Liberles D. 2011. Evolution after gene duplication. Hoboken (NJ): John Wiley & Sons.
Domman DB, Steven BT, Ward NL. 2011. Random transposon mutagenesis of *Verrucomicrobium spinosum* DSM 4136(T). Arch Microbiol. 193:307–312.
Edgar R. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32:1792–1797.
Felsenstein J. 1989. PHYLIP—phylogeny inference package (version 3.2). Cladistics 5:164–166.
Fieseler L, Horn M, Wagner M, Hentschel U. 2004. Discovery of the novel candidate phylum "Poribacteria" in marine sponges. Appl Environ Microbiol. 70:3724–3732.
Fitzpatrick D, Logue ME, Stajich JE, Butler G. 2006. A fungal phylogeny based on 42 complete genomes derived from supertree and combined gene analysis. BMC Evol Biol. 6:99.
Forslund K, Henricson A, Hollich V, Sonnhammer E. 2008. Domain tree-based analysis of protein architecture evolution. Mol Biol Evol. 25:254–264.
Franzmann P, Skerman V. 1984. *Gemmata obscuriglobus*, a new genus and species of the budding bacteria. Antonie van Leeuwenhoek 50: 261–268.
Fraser C, Hanage WP, Spratt BG. 2007. Recombination and the nature of bacterial speciation. Science 315:476–480.
Fritsche T, Sobek D, Gautom R. 1998. Enhancement of in vitro cytopathogenicity by *Acanthamoeba* spp. following acquisition of bacterial endosymbionts. FEMS Microbiol Lett. 166:231–236.
Fuerst JA. 2005. Intracellular compartmentation in *Planctomycetes*. Annu Rev Microbiol. 59:299–328.
Gautom R, Fritsche T. 1995. Transmissibility of bacterial endosymbionts between isolates of *Acanthamoeba* spp. J Eukaryot Microbiol. 42: 452–456.
Goldstein E, Citron DM, Peraino VA, Cross SA. 2003. *Desulfovibrio desulfuricans* bacteremia and review of human *Desulfovibrio* infections. J Clin Microbiol. 41:2752–2754.
Gu X. 1999. Statistical methods for testing functional divergence after gene duplication. Mol Biol Evol. 16:1664–1674.
Guindon S, Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol. 52: 696–704.
Hanage W, Fraser C, Spratt B. 2005. Fuzzy species among recombinogenic bacteria. BMC Biol. 3:6.
Horn M. 2008. *Chlamydiae* as symbionts in eukaryotes. Annu Rev Microbiol. 62:113–131.
Hou S, et al. 2008. Complete genome sequence of the extremely acidophilic methanotroph isolate V4, *Methylacidiphilum infernorum*, a representative of the bacterial phylum *Verrucomicrobia*. Biol Direct. 3:26.
Huson D, Bryant D. 2006. Application of phylogenetic networks in evolutionary studies. Mol Biol Evol. 23:254–267.
Innan H, Kondrashov F. 2010. The evolution of gene duplications: classifying and distinguishing between models. Nat Rev Genet. 11:97–108.
Jain R, Rivera M, Moore J, Lake J. 2003. Horizontal gene transfer accelerates genome innovation and evolution. Mol Biol Evol. 20: 1598–1602.
Jogler C, Glöckner F, Kolter R. 2011. Characterization of *Planctomyces limnophilus* and development of genetic tools for its manipulation establish it as a model species for the phylum *Planctomycetes*. Appl Environ Microbiol. 77:5826–5829.
Kalman S, et al. 1999. Comparative genomes of *Chlamydia pneumoniae* and *C. trachomatis*. Nat Genet. 21:385–389.
Kamneva O, Liberles D, Ward N. 2010. Genome-wide influence of indel substitutions on evolution of bacteria of the PVC super-phylum, revealed using a novel computational method. Genome Biol Evol. 2: 870–886.
Kanehisa M, Goto S, Furumichi M, Tanabe M, Hirakawa M. 2010. KEGG for representation and analysis of molecular networks involving diseases and drugs. Nucleic Acids Res. 38:D355–D360.
Kelly G, et al. 1999. Structure of the cell-adhesion fragment of intimin from enteropathogenic *Escherichia coli*. Nat Struct Biol. 6:313–318.
Knudsen B, Miyamoto M. 2001. A likelihood ratio test for evolutionary rate shifts and functional divergence among proteins. Proc Natl Acad Sci U S A. 98:14512–14517.
Lartillot N, Lepage T, Blanquart S. 2009. PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. Bioinformatics 25:2286–2288.
Lee P, Tullman-Ercek D, Georgiou G. 2006. The bacterial twin-arginine translocation pathway. Annu Rev Microbiol. 60:373–395.
Lee K, et al. 2009. Phylum *Verrucomicrobia* representatives share a compartmentalized cell plan with members of bacterial phylum *Planctomycetes*. BMC Microbiol. 9:5.
Lefébure T, Stanhope M. 2007. Evolution of the core and pan-genome of *Streptococcus*: positive selection, recombination, and genome composition. Genome Biol. 8:R71–R71.
Letunic I, Bork P. 2011. Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy. Nucleic Acids Res. 39: W475–W478.
Li L, Stoeckert C, Roos D. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. 13:2178–2189.
McDonald DB, Parchman TL, Bower MR, Hubert WA, Rahel FJ. 2008. An introduced and a native vertebrate hybridize to form a genetic bridge to a second native species. Proc Natl Acad Sci U S A. 105: 10837–10842.
McInerney J, Pisani D, Bapteste D, O'Connell M. 2011. The public goods hypothesis for the evolution of life on Earth. Biol Direct. 6:41.
Mira A, Ochman H, Moran N. 2001. Deletional bias and the evolution of bacterial genomes. Trends Genet. 17:589–596.
Nakjang S, Ndeh DA, Wipat A, Bolam DN, Hirt RP. 2012. A novel extracellular metallopeptidase domain shared by animal host-associated mutualistic and pathogenic microbes. PLoS One 7:e30287.
Ochman H, Lawrence J, Groisman E. 2000. Lateral gene transfer and the nature of bacterial innovation. Nature 405:299–304.
Pagani I, et al. 2012. The Genomes OnLine Database (GOLD) v.4: status of genomic and metagenomic projects and their associated metadata. Nucleic Acids Res. 40:D571–D579.
Pagel M. 1999. The maximum likelihood approach to reconstructing ancestral character states of discrete characters on phylogenies. Syst Biol. 48:612–622.
Pilhofer M, et al. 2008. Characterization and evolution of cell division and cell wall synthesis genes in the bacterial phyla *Verrucomicrobia*, Lentisphaerae, *Chlamydiae*, and *Planctomycetes* and phylogenetic comparison with rRNA genes. J Bacteriol. 190:3192–3202.

Reeves J. 1992. Heterogeneity in the substitution process of amino acid sites of proteins coded for by mitochondrial DNA. J Mol Evol. 35:17–31.
Retchless A, Lawrence J. 2007. Temporal fragmentation of speciation in bacteria. Science 317:1093–1096.
Retchless A, Lawrence J. 2010. Phylogenetic incongruence arising from fragmented speciation in enteric bacteria. Proc Natl Acad Sci U S A. 107:11453–11458.
Rivera M, Jain R, Moore J, Lake J. 1998. Genomic evidence for two functionally distinct gene classes. Proc Natl Acad Sci U S A. 95:6239–6244.
Roberton A, Stanley R. 1982. In vitro utilization of mucin by *Bacteroides fragilis*. Appl Environ Microbiol. 43:325–330.
Sait M, et al. 2011. Genomic and experimental evidence suggests that *Verrucomicrobium spinosum* interacts with eukaryotes. Front Microbiol. 2:211.
Santarella-Mellwig R, et al. 2010. The compartmentalized bacteria of the *Planctomycetes-Verrucomicrobia-Chlamydiae* superphylum have membrane coat-like proteins. PLoS Biol. 8:e1000281.
Sharon N, Lis H. 1972. Lectins: cell-agglutinating and sugar-specific proteins. Science 177:949–959.
Smillie C, et al. 2011. Ecology drives a global network of gene exchange connecting the human microbiome. Nature 480:241–244.
Studholme D, Fuerst JA, Bateman A. 2004. Novel protein domains and motifs in the marine planctomycete *Rhodopirellula baltica*. FEMS Microbiol Lett. 236:333–340.
Sullivan M, et al. 2006. Prevalence and evolution of core photosystem II genes in marine cyanobacterial viruses and their hosts. PLoS Biol. 4: e234.
Turroni F, et al. 2010. Genome analysis of *Bifidobacterium Bifidum* PRL2010 reveals metabolic pathways for host-derived glycan foraging. Proc Natl Acad Sci U S A. 107:19514–19519.
Wagner M, Horn M. 2006. The *Planctomycetes*, *Verrucomicrobia*, *Chlamydiae* and sister phyla comprise a superphylum with biotechnological and medical relevance. Curr Opinion Biotechnol. 17:241–249.
Whelan S, Goldman N. 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol Biol Evol. 18:691–699.
Xu G, Min J. 2011. Structure and function of WD40 domain proteins. Protein Cell 2:202–214.
Yang Z. 1993. Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. Mol Biol Evol. 10: 1396–1401.
Yeats C, Bentley S, Bateman A. 2003. New knowledge from old: in silico discovery of novel protein domains in *Streptomyces coelicolor*. BMC Microbiol. 3:3.

Associate editor: Bill Martin