Genome-Wide Influence of Indel Substitutions on Evolution of Bacteria of the PVC Superphylum, Revealed Using a Novel Computational Method

Olga K. Kamneva[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: 29 October 2010

Abstract

Whole-genome scans for positive Darwinian selection are widely used to detect evolution of genome novelty. Most approaches are based on evaluation of nonsynonymous to synonymous substitution rate ratio across evolutionary lineages. These methods are sensitive to saturation of synonymous sites and thus cannot be used to study evolution of distantly related organisms. In contrast, indels occur less frequently than amino acid replacements, accumulate more slowly, and can be employed to characterize evolution of diverged organisms. As indels are also subject to the forces of natural selection, they can generate functional changes through positive selection. Here, we present a new computational approach to detect selective constraints on indel substitutions at the whole-genome level for distantly related organisms. Our method is based on ancestral sequence reconstruction, takes into account the varying susceptibility of different types of secondary structure to indels, and according to simulation studies is conservative. We applied this newly developed framework to characterize the evolution of organisms of the *Planctomycetes*, *Verrucomicrobia*, *Chlamydiae* (PVC) bacterial superphylum. The superphylum contains organisms with unique cell biology, physiology, and diverse lifestyles. It includes bacteria with simple cell organization and more complex eukaryote-like compartmentalization. Lifestyles range from free-living organisms to obligate pathogens. In this study, we conduct a whole-genome level analysis of indel substitutions specific to evolutionary lineages of the PVC superphylum and found that indels evolved under positive selection on up to 12% of gene tree branches. We also analyzed possible functional consequences for several case studies of predicted indel events.

Key words: selection, indel substitutions, PVC superphylum.

Introduction

Genome-wide scans for positive Darwinian selection are widely used to detect genetic changes underlying, or associated with, emergence of novel traits (Liberles et al. 2001; Davids et al. 2002; Clark et al. 2003; Lefébure and Stanhope 2007, 2009; Petersen et al. 2007; Orsi et al. 2008). The majority of methods for this kind of analysis rely on evaluation of nonsynonymous to synonymous substitutions rate ratio (K_a/K_s) across different lineages (Messier and Stewart 1997; Benner et al. 1998; Liberles 2001; Yang and Nielsen 2002; Zhang et al. 2005). These approaches are sensitive to saturation of synonymous sites (Smith JM and Smith NH 1996), and therefore most suitable for use on data sets containing closely related sequences (Anisimova and Liberles 2007). For this reason, studies applying such techniques to bacterial sequences have mostly analyzed evolution of strains of one species (Davids et al. 2002; Lefébure and Stanhope 2007, 2009) or species of one genus (Petersen et al. 2007; Orsi et al. 2008), where it is possible to correctly estimate K_s and K_a values. Another problem associated with this approach is selection on synonymous sites, although recently published methodology may enable detection of, and correction for, this phenomenon (Zhou et al. 2010).

Large evolutionary distances between sequences, observed when bacteria are related at higher taxonomic levels such as the phylum, require analysis at the protein level. There are several methods developed for evolutionary analysis at this level (reviewed in Anisimova and Liberles [2007]). Many methods compare evolutionary rates of amino acid substitutions across different phylogenetic lineages (Gu et al. 1995; Gu 1999, 2001; Knudsen and Miyamoto 2001; Pupko and Galtier 2002; Blouin et al. 2003; Abhiman and Sonnhammer 2005; Dorman 2007; Penn et al. 2008). These approaches aim to detect overall changes in selective constraints on individual amino acid sites over time, allowing inference of changes in the biological function of the protein. Such changes are detected through one of two signals, consistent with changes in rates at a site (rate-shifting sites or type I functional divergence) or changes in conservation at a site (conservation-shifting sites or type II functional divergence). However, the limited statistical power of the most recently published and rigorous methods does not allow testing for evolutionary rate shifts on small or evolutionarily sparse data sets.

Indel substitutions represent another common type of sequence variation contributing to the evolution of both coding and regulatory/noncoding sequences (Britten 2002; Osterberg et al. 2002; Britten et al. 2003; Podlaha and Zhang 2003; Podlaha et al. 2005; Schully and Hellberg 2006; Brandstrom and Ellegren 2007; Chan et al. 2007; Chen et al. 2010). However, rates and patterns of indels are less well studied than those associated with nucleotide or amino acid replacement substitutions. Beyond the rates and patterns of indels, further considerations include the evolutionary dynamics and functional consequences of change, which are even less well studied. Lastly, almost all types of evolutionary sequence analysis ignore functional information in gaps.

The underlying mechanisms driving the insertion or deletion of protein segments remain unclear. Insertions and deletion are expected to occur through recombination, polymerase error, and DNA repair enzyme error in individuals within a population. As with point mutation, there is interplay between selection and other factors, resulting in the differential fixation of indels. Empirical studies have shown that indels are more easily accommodated within coiled regions of a protein evolving under evolutionary constraints (Benner and Gerloff 1991). However, there are several known examples when length of the loop affects function of the protein by altering affinity of protein–protein interactions (Meenan et al. 2010) or changing thermodynamics of proper folding (Viguera and Serrano 1997). The indel length distribution is best described by a Zipfian distribution rather than an exponential distribution (Chang and Benner 2004). It has also been shown that indels accumulate more slowly than amino acid substitutions. The stochastic process of indel accumulation reaches saturation at an amino acid identity level of about 15% (Pascarella and Argos 1992). Synonymous substitutions in DNA show saturation at 67% protein sequence identity, assuming a K_a/K_s ratio of 0.2 and an upper measurement bound of $K_s = 2$, using a Poisson correction for multiple hits and an assumption of equal rates across sites (Yang 2006). A nonhomogeneous distribution of indels across genomes and/or evolutionary time has been identified for several species (Brandstrom and Ellegren 2007; Chen et al. 2010). This can be interpreted as a sign of selection (positive or purifying) or of varying mutational opportunities for insertion and deletion. In case studies of individual genes, including the *Catsper1* calcium ion channel genes in mammals (Podlaha and Zhang 2003; Podlaha et al. 2005) and the Acp26Aa gene in *Drosophila* species (Schully and Hellberg 2006), it has been found that positive diversifying selection acts upon indels. Here, we have expanded on this general concept in a genome-wide study that creates a new approach for evolutionary analysis. The low accumulation rate of indel substitutions makes this kind of methodology appropriate for investigation of the evolution of distantly related organisms and allows insight into the evolutionary role of this less well-understood type of sequence variation. We applied the developed method to reveal the role of indel substitution in the evolution of gene families constructed for the *Planctomycetes*, *Verrucomicrobia*, *Chlamydiae* (PVC) superphylum.

The PVC superphylum is currently described as a group of four bacterial phyla, *Planctomycetes*, *Verrucomicrobia*, *Chlamydiae*, and *Lentisphaerae* (together with *Poribacteria* and *OP3* candidate phyla). The relative phylogenetic positions of the organisms belonging to these four phyla have been debated (Stackebrandt et al. 1984; Roenner et al. 1991; Embley et al. 1994; Van de Peer et al. 1994; Hedlund et al. 1997; Ward et al. 2000, 2006; Jenkins and Fuerst 2001; Schloss and Handelsman 2004; Wagner and Horn 2006; Griffiths and Gupta 2007; Pilhofer et al. 2008). Initially planctomycetes were considered to be a deep-branching bacterial lineage (Stackebrandt et al. 1984; Roenner et al. 1991). This hypothesis was later rejected based on analysis of larger data sets with more sophisticated methods. Some early studies suggested chlamydiae to be the closest relative of planctomycetes (Embley et al. 1994; Van de Peer et al. 1994); later it was shown that verrucomicrobia are the closest living relatives of chlamydiae (Griffiths and Gupta 2007). More recently, it was established that the four phyla form a coherent group of organisms (Schloss and Handelsman 2004; Wagner and Horn 2006; Pilhofer et al. 2008). Thus, the clear evolutionary relationships between phyla in this superphylum provide an opportunity to define the root of the species tree and, consequently, place the root on every gene phylogeny. This step is essential in studying the process of insertion and deletions.

The organisms of the superphylum are physiologically divergent. They include obligate human pathogens, free-living soil and aquatic microorganisms, and organisms found in close association with metazoan hosts (Wagner and Horn 2006; Ward et al. 2006). Differences in the population structure associated with these lifestyles not only give us a chance

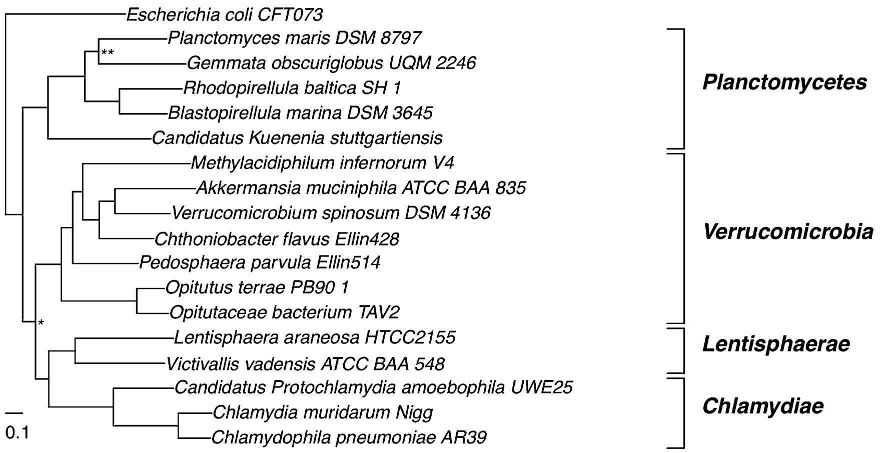

Escherichia coli CFT073
Planctomyces maris DSM 8797
**
Gemmata obscuriglobus UQM 2246
Rhodopirellula baltica SH 1
Blastopirellula marina DSM 3645
Candidatus Kuenenia stuttgartiensis
Methylacidiphilum infernorum V4
Akkermansia muciniphila ATCC BAA 835
Verrucomicrobium spinosum DSM 4136
Chthoniobacter flavus Ellin428
Pedosphaera parvula Ellin514
Opitutus terrae PB90 1
Opitutaceae bacterium TAV2
*
Lentisphaera araneosa HTCC2155
Victivallis vadensis ATCC BAA 548
Candidatus Protochlamydia amoebophila UWE25
Chlamydia muridarum Nigg
Chlamydophila pneumoniae AR39
0.1
Planctomycetes
Verrucomicrobia
Lentisphaerae
Chlamydiae

Fig. 1.—PVC species tree. Maximum likelihood phylogeny was reconstructed using concatenated phylogenetic markers, with *Escherichia coli* CFT073 serving as an outgroup. Bootstrap support values of 100% are not shown, clades marked with * and ** have 91% and 68% support, respectively (percentage of 1,000 nonparametric bootstrap runs). Phyla names are shown on the right.

to study the emergence of new lifestyles but also allow testing of our newly developed method because nonrandom processes (selection) should have larger effects on organisms with large effective population sizes.

All the characterized planctomycetes, verrucomicrobia, and the recently discovered poribacteria have a common cell plan that features an additional intracellular membrane and is unique among the bacteria (Lindsay et al. 2001; Fieseler et al. 2004; Fuerst 2005; Lee et al. 2009). Planctomycete bacteria exhibit variations from this common plan (Fuerst 2005). Based on protein structural analysis and immunomicroscopy, it has been suggested that the cell compartmentalization machinery in these bacteria is similar to that utilized by eukaryotes, raising the possibility that organisms ancestral to the PVC superphylum contributed to eukaryogenesis (Santarella-Mellwig et al. 2010). Endocytosis (a stereotypically eukaryotic trait) has recently been demonstrated in the planctomycete *Gemmata obscuriglobus* (Lonhienne et al. 2010). Many species of the superphylum also develop unique extracellular structures, such as prosthecae, stalks, polar holdfasts, and fimbriae (Hedlund et al. 1997; Cho et al. 2004; Wagner and Horn 2006; Ward et al. 2006). Therefore, study of this group may provide information on the evolution of complex intracellular and extracellular structures in bacteria and lineage-specific processes associated with the development of host association and acquisition of pathogenic potential. Genome sequences for a number of organisms of the PVC superphylum have recently become available (evolutionary relationships of the species with available genome sequences are depicted in fig. 1). The availability of genomic information permits study of the evolutionary history of this remarkable group of bacteria.

Thus, we are interested in revealing the general patterns and evolutionary consequences of indel substitutions, as well as their role in the evolution of organisms of the PVC superphylum. In the present study, we report a systematic analysis of indels specific for different evolutionary lineages of the superphylum, their association with molecular structure and function, and their possible functional effects.

Materials and Methods

Sequence Data

Protein sequences for every organism of the PVC superphylum with an available genome sequence (fig. 1) were downloaded from GenBank. The final list of analyzed species included (NCBI Taxonomy ID is shown in parentheses) *G. obscuriglobus* UQM 2246^T (214688), *Planctomyces maris* DSM 8797^T (344747), *Rhodopirellula baltica* SH 1^T (243090), *Blastopirellula marina* DSM 3645^T (314230), *Candidatus* Kuenenia stuttgartiensis (174633), *Opitutaceae* bacterium TAV2 (278957), *Opitutus terrae* PB90 1^T (452637), *Pedosphaera parvula* Ellin514^T (320771), *Chthoniobacter flavus* Ellin428^T (497964), *Akkermansia muciniphila* ATCC BAA 835^T (349741), *Verrucomicrobium spinosum* DSM 4136^T (240016), *Methylacidiphilum infernorum* V4^T (481448), *Chlamydia muridarum* Nigg^T (243161), *Chlamydophila pneumoniae* AR39^T (115711), *Candidatus* Protochlamydia amoebophila UWE25^T (264201), *Victivallis vadensis* ATCC BAA 548^T (340101), and *Lentisphaera araneosa* HTCC2155^T (313628). To incorporate compatible evolutionary distances, only one representative of each genus (three species) from the phylum *Chlamydiae* was included in the analysis.

Gene Families

Gene families were identified using the OrthoMCL software (Li et al. 2003) with an inflation value of 1.5 and the threshold for expectation value for Blast search (Altschul et al. 1990) set at 0.00001. Protein families obtained with OrthoMCL were subsequently refined using 25% identity and 70% of alignment extension thresholds for complete linkage clustering based upon pairwise alignments using MUSCLE (Edgar 2004).

Sequence Alignments and Gene Phylogeny

Sequences from refined gene families containing four or more members were aligned using MUSCLE (Edgar 2004). Alignment quality was assessed using average parsimony score (Fitch 1971) in sliding window analysis, where we compared the average score of the alignment obtained with the average score of a randomized alignment with gap retention. Regions with lower than expected parsimony score were excluded from further consideration.

Phylogeny for every gene family was reconstructed using the RAxML software (Stamatakis 2006) implementing the WAG + I + GAMMA + F (Reeves 1992; Yang 1993; Whelan and Goldman 2001) evolutionary model, as this model was found to be the model of best fit for the concatenated alignment used for species tree reconstruction as well as for the large majority of individual gene tree alignments. Gene trees were rooted using SoftParsMap (Berglund-Sonnhammer et al. 2006), and a species tree calculated as described below.

DNA-Based Analysis and Evolutionary Rate-Shift Analysis

We employed the yn00 program from the PAML 4.1 package to check pairwise distances (in number of synonymous substitutions per codon of alignment) between sequences in the gene families. DNA alignments were obtained using DNA sequences, based on protein alignments for the gene families. Columns of DNA alignments (in codons) corresponding to the filtered-out columns of protein alignments were also excluded from the final alignments. Pairwise distances were obtained for every pair of sequences in the gene family for every gene family. The average value of pairwise K_s distance was calculated.

For rate-shift analysis, RASER Version 1.1 (Penn et al. 2008) was run twice (once for the rate-shift enabling model and once for the null model). Log-likelihood values for both models were compared using the likelihood ratio test.

Species Tree Reconstruction

In order to reconstruct a species tree, we extracted gene families containing exactly one sequence from every species under consideration; there were 53 such gene families. Homologous sequences from the *Escherichia coli* CFT073 genome were added to every extracted gene family when it was possible to obtain exactly one related sequence; otherwise the gene family was excluded from the species tree reconstruction. Two gene families were discarded on this basis. Individual gene families were also evaluated based on the robustness of the resulting phylogenetic signal, determined by performing 50 bootstrap runs for each gene family. Due to an average support value for the consensus tree of less than 50%, one additional gene family was excluded from the species tree reconstruction. Thus, 50 gene families were included in the species tree reconstruction. The evolutionary model of best fit was determined for every gene family in a set of 51 using ProtTest (Abascal et al. 2005). We found that four general models of evolution were supported by the gene families: Whelan and Goldman model (WAG)—48; Blosum62—1; Dayhoff—1; Jones, Taylor, and Thorton (JTT)—1 gene families (Dayhoff and Schwartz 1978; Henikoff S and Henikoff JG 1992; Jones et al. 1992; Whelan and Goldman 2001). Alignments for individual gene families supporting variations of the WAG model (with addition of rates across sites variation and empirically estimated base frequencies) were concatenated, and phylogeny was determined using RAxML (Stamatakis 2006) implementing the WAG + GAMMA + F evolutionary model (Reeves 1992; Yang 1993; Whelan and Goldman 2001). The three gene families supporting JTT, Blosum62, and Dayhoff evolutionary models were not considered further. Initially, 50 trees were reconstructed for the concatenated set, then the best tree was chosen based on likelihood value. These trees were subsequently tested with 250 nonparametric bootstraps performed using the same evolutionary models in RAxML (Stamatakis 2006). The resulting trees were rooted using *E. coli* CFT073 as an outgroup.

The concatenated alignment used for species tree reconstruction was also tested for support of the phylogenetic network, using the NeighbourNet algorithm implemented in the SplitsTree software package (Huson and Bryant 2006).

Branch-Specific Insertion/Deletion Rate Estimation for Different Types of Secondary Structure

We used ancestral sequence reconstruction to obtain raw data on insertion/deletion rates. Gapped ancestral sequence reconstruction was performed on the full-length alignments using GASP (Edwards and Shields 2004) and obtained as described above, from alignments and phylogenetic trees for individual gene families. Secondary structure for one representative extant sequence from each gene family was predicted using PSIPRED (Jones 1999). The alignment for each gene family (including both extant sequences and predicted ancestral sequences) was split into three subalignments based on the assigned type of secondary structure (alpha-helices, beta-strands, and coils) for the representative gene family member. Parts of the same secondary structure longer than six amino acids were concatenated. In this way, we split the alignment of every gene family into three parts and examined indel patterns on every partition as well as on initial full-length alignments. Branch lengths for every gene tree for every secondary structure type were reevaluated using the ProtDist program within the PHYLIP package (Felsenstein 1989). The most parsimonious number of indels along branches of the gene phylogeny was inferred by comparison of the descendant state relative with the ancestral state.

Insertions or deletions of longer sequence elements are expected to have stronger propensities for functional consequences in the protein. Therefore, indels were analyzed in four different groups: all indels, indels of at least two amino acids, at least three amino acids, and at least four amino acids. Observed branch-specific insertion/deletion rates as number of indels per unit of evolutionary distance per unit (one substitution per amino acid site) of alignment length (equivalent of 1,000 bp of DNA sequence) were determined for every branch of the gene phylogeny for every gene family as:

$$\boldsymbol{R}_{ij}^{o} = \frac{(\boldsymbol{N}_{ij}^{o} \times 333)}{\boldsymbol{A}_{ij} \times \boldsymbol{B}_{ij}},$$

where: $\boldsymbol{R}_{ij}^{o}$, observed insertion/deletion rate for branch i from phylogenetic tree j; $\boldsymbol{N}_{ij}^{o}$, observed number of insertions/deletions that occurred on branch i of phylogenetic tree j; $\boldsymbol{A}_{ij}$, alignment length of branch i from phylogenetic tree j; $\boldsymbol{B}_{ij}$, branch length of i from phylogenetic tree j; 333, an equivalent of 1,000 bp of DNA sequence.

Division of the number of events by branch length allows normalization of this number per unit of evolutionary distance between ancestral and descendant sequences. Division by alignment length and subsequent multiplication by 333 allows normalization of the number of events per unit of alignment length corresponding to 1,000 bp of DNA alignment.

Generation of Expected Insertion/Deletion Rate Distribution

To obtain expected values of insertion/deletion rates, we performed simulations for different types of secondary structure, as well as for original full-length proteins. Insertion/deletion events were randomly assigned to the branches of gene trees. The probability of the insertion/deletion being assigned to a particular branch was proportional to the branch and alignment lengths and defined as:

$$\boldsymbol{P}_{ij} = \frac{(\boldsymbol{A}_{ij} \times \boldsymbol{B}_{ij})}{\sum\limits_{i=1,j=1}^{i=s,j=n} (\boldsymbol{A}_{ij} \times \boldsymbol{B}_{ij})},$$

where: $\boldsymbol{P}_{ij}$, probability of insertion/deletion to occur on branch i; $\boldsymbol{A}_{ij}$, alignment length of branch i from phylogenetic tree j; $\boldsymbol{B}_{ij}$, branch length of i from phylogenetic tree j; n, number of gene families; s, number of branches in phylogenetic tree j.

The final count of assigned events was equivalent to the observed number of insertions/deletions in the corresponding data set. Expected insertion/deletion rates were determined as described above for observed rates. The distributions of insertion/deletion rates varied depending on the type of secondary structure and length of insertions or deletion. Therefore, false discovery rate (FDR) was employed to identify the appropriate confidence level for determining significant deviations of observed data from random process. Individual indel rate values corresponding to 50% FDR were determined for every type of secondary structure as well as for full-length alignments, for every group of indel length. This FDR was used as a threshold value in the further identification of the branches of gene trees with significantly elevated insertion/deletion rates.

Performance Evaluation

In order to evaluate the performance of the ancestral sequence reconstruction methodology in enabling detection of the branches where indels occurred, 12 artificial data sets were simulated as follows. 1) 10% of branches from the entire data set were randomly assigned to have indels under positive selection (foreground branches). 2) The probability of the event being assigned to a particular branch was proportional to the branch, to the alignment length, and to the scaling factor and was defined as:

$$\boldsymbol{P}_{ij}^{m} = \frac{(\boldsymbol{A}_{ij} \times \boldsymbol{B}_{ij} \times \boldsymbol{x}_{ij})}{\sum\limits_{i=1,j=1}^{i=s,j=n} (\boldsymbol{A}_{ij} \times \boldsymbol{B}_{ij} \times \boldsymbol{x}_{ij})},$$

where: $\boldsymbol{P}_{ij}$, probability of insertion/deletion to occur on branch i; $\boldsymbol{A}_{ij}$, alignment length of branch i from phylogenetic tree j; $\boldsymbol{B}_{ij}$, branch length of i from phylogenetic tree j; n, number of gene families; s, number of branches in phylogenetic tree j; $\boldsymbol{x}_{ijm}$, scaling factor derived from a gamma distribution with shape k and scale θ; m, rate acceleration mode on foreground branches.

Although alignment length and branch length were always the same for a particular branch i of a particular gene tree j, parameters of the distribution from which scaling factor is sampled varied, depending on if the branch under consideration was background or foreground and on the rate acceleration mode (m) on the foreground branch.

$m = 1$:

$$\boldsymbol{x}_{ij} \sim \begin{cases} \Gamma(k = 1, \theta = 1) + 1, \text{if } i \text{ is a background branch of tree } j, \\ \Gamma(k = 1, \theta = 1) + 1, \text{if } i \text{ is a foreground branch of tree } j. \end{cases}$$

$m = 2$:

$$\boldsymbol{x}_{ij} \sim \begin{cases} \Gamma(k=1,\theta=1) + 1, \text{if } i \text{ is a background branch of tree } j, \\ \Gamma(k=0.5,\theta=2,000) + 5, \text{if } i \text{ is a foreground branch of tree } j. \end{cases}$$

$m = 3$:

$$\boldsymbol{x}_{ij} \sim \begin{cases} \Gamma(k=1,\theta=1) + 1, \text{if } i \text{ is a background branch of tree } j, \\ \Gamma(k=0.5,\theta=2,000) + 10, \text{if } i \text{ is a foreground branch of tree } j. \end{cases}$$

$m = 4$:

$$\boldsymbol{x}_{ij} \sim \begin{cases} \Gamma(k=1,\theta=1) + 1, \text{if } i \text{ is a background branch of tree } j, \\ \Gamma(k=0.5,\theta=2,000) + 100, \text{if } i \text{ is a foreground branch of tree } j. \end{cases}$$

Thus, *m* equals to 1 corresponds to neutral evolution of indel substitutions on foreground branches and should serve as a negative control (no branches should be detected to have indels under positive selection). *m* equals 2, 3, or 4 corresponds to evolutionary scenarios with positive selection of different strengths on indel substitutions on foreground branches. Finally, 100,000, 10,000, or 1,000 events were assigned to branches of gene trees. We used different number of events in order to assess the influence of absolute counts of events on the performance of the algorithm, as we observed very different number of insertions/deletion in different types of secondary structure. Observed counts of assigned events were treated as real counts of insertions/deletions, where corresponding rate values as well as the expected distributions of event rates were calculated as described above. For all 12 simulated data sets, the sensitivity and specificity were determined for threshold values ranging from –0.01 to the maximum rate +1 to derive receiver operating characteristic (ROC) curves. Finally, rates and confidence values corresponding to 50% FDR (if it was possible) were determined for every simulated data set and percentage of false positive (FP) and false negative (FN) were calculated.

Mapping onto Metabolic Pathways, Structural Modeling, Data Management

In order to link the 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 (Kanehisa et al. 2010). Gene families were assigned to pathways based on 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 in order to determine statistical overrepresentation of cellular pathways among the gene families where positive selection on indels was determined. We transformed the *P* values obtained for every pathway to generate heatmap charts based on these transformed values. The transformation scaled all *P* values on a scale from −1 to 1, where values approaching −1 corresponded to underrepresented pathways, and values approaching 1 corresponded to overrepresented pathways.

Modeling of the 3D structure of ribosomal protein L17 from *G. obscuriglobus* was performed using the i-TASSER web server (Roy et al. 2010) with template structure 2WRJ:R. It allowed prediction of the structure of inserted fragments that did not have a homologous region in any template. Three-dimensional structures of the proteins were visualized using PyMOL (DeLano 2002).

Data manipulations were performed using custom R and Perl scripts.

Results and Discussion

To investigate the patterns of selective constraints on indel substitutions in a genome-wide manner, we estimated secondary structure-specific insertion and deletion rates for every lineage of every gene family in the data set. We compared the observed distribution of insertion/deletion rates with the expected distributions obtained using simulations under neutral conditions. We applied this approach to a data set of 17 genomes from members of the PVC superphylum to evaluate how insertions and deletions have affected the evolution of this group of distantly related organisms.

PVC Species Tree

The main subject of our study was the process of insertions and deletions of DNA fragments in predicted protein-coding sequences, analyzed through comparison of ancestral and descendant states. Thus, in order to differentiate between insertions and deletions, we needed to define ancestor-descendant relationships between reconstructed ancestral sequences. This was possible only through analysis of a rooted phylogeny. We chose to use a species tree to place a root on the gene tree for every analyzed gene family. Alternatively, we could have used outgroup rooting, but generating gene families using outgroup species and constructing reliable alignments and phylogenies would have been problematic.

We reconstructed a species tree using all 14 available genome sequences of organisms from phyla *Planctomycetes*, *Lentisphaera*, and *Verrucomicrobia*. In order to

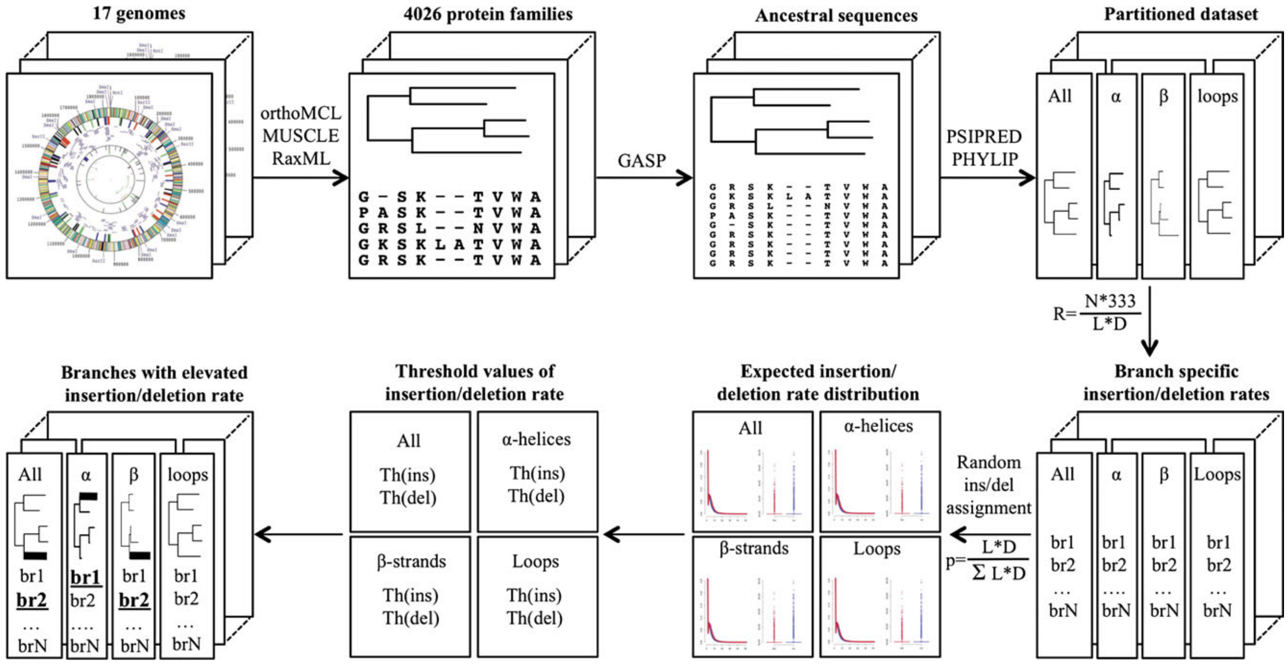


Fig. 2.—Genome-wide indel analysis pipeline.

include compatible evolutionary distances, only one representative of every genus (three species) from the phylum *Chlamydiae* was included in the analysis. The phylogenetic relationships of a total 17 species representing four phyla in the PVC superphylum were determined using the maximum likelihood method (fig. 1). Phylogeny was derived using concatenated phylogenetic markers that had clear one-to-one homologs in every genome under consideration and in *E. coli* CFT073 and supported the same model of protein evolution. They included mostly informational genes considered to be resistant to horizontal gene transfer (Rivera et al. 1998; Sorek et al. 2007). A test for signal of a network-like structure in the concatenated alignment used for species tree reconstruction recovered some signal of a phylogenetic network at the root of the tree (supplementary fig. 1, Supplementary Material online), which is consistent with the fragmented nature of speciation and loss of phylogenetic signal observed for other bacteria (Retchless and Lawrence 2010). The obtained species tree (fig. 1) was largely consistent with most recently published 16S- and 23S rRNA-based phylogenies (Wagner and Horn 2006; Pilhofer et al. 2008). Species of four distinct phyla formed four monophyletic groups. Planctomycetes occupied a separate position from the rest of the superphylum, and *Kuenenia stuttgartiensis* appeared to be the most ancestral lineage among planctomycetes, as was observed in previous studies. Within the rest of the superphylum, *Lentisphaera* formed a cluster with *Chlamydiae*, which contradicts previously published phylogenies where *Lentisphaera* species were more closely related to phylum *Verrucomicrobia* (Wagner and Horn 2006; Pilhofer et al. 2008).

Genome-Wide Characterization of Indel Substitutions

We included 17 complete (finished or draft) genomes for 17 representative species of the PVC superphylum with a total of 78,728 protein sequences used for subsequent analysis (fig. 2). Markov clustering and subsequent filtering based on sequence identity and alignment extension allowed us to define 43,960 homologous families (supplementary fig. 2, Supplementary Material online), including 3,959 gene families containing four or more sequences. Patterns of indel substitutions were studied using a newly developed approach. Every gene tree was rooted using either midpoint rooting (if the gene family consisted of sequences from one organism) or most parsimonious gene tree/species tree reconciliation based upon a minimization of the number of inferred duplication events with different root placements (Berglund-Sonnhammer et al. 2006).

In the next step, ancestral sequences were reconstructed for every node of every gene tree for all the obtained gene families of appropriate size (containing four or more sequences), using gapped ancestral sequence reconstruction. This approach merges the maximum likelihood method of ancestral sequence reconstruction for substitutions, with the most parsimonious assignment of insertions and deletion to the branches of the phylogeny (Edwards and Shields 2004). Comparison of ancestral and descendant sequences permitted assessment of lineage-specific insertions and deletions for each branch of every gene family. Known branch length and alignment length allowed inference of the branch-specific insertion/deletion rate for every branch of every gene family. This new approach for studying the evolution of insertions/deletions allows for greater sensitivity, as it permits testing for significant deviation from the neutral expectation on individual branches of a gene tree rather than simply evaluating events using extant sequences and pairwise analysis.

In order to determine whether the stochastic process of accumulation of indel substitutions has reached saturation at the observed level of evolutionary divergence, we examined the number of insertions and deletions per unit of evolutionary distance (supplementary fig. 3, Supplementary Material online). It is clear that the number of indels exhibits a linear relationship with evolutionary distances between sequences (insertions $R = 0.1$, deletions $R = 0.15$). This was also supported by combined indel data from ancestral sequence reconstruction as well as by the data derived from alignments alone (alignments $R = 0.23$, ancestral sequences $R = 0.33$).

Secondary Structure-Specific Insertion/Deletion Rates

It has been previously shown that different types of secondary structure have different susceptibility to insertion and deletion (Benner and Gerloff 1991). Loops or coils accommodate insertions or deletions more easily than alpha-helices or beta-strands. To evaluate secondary structure-specific patterns of indel substitutions, alignments in every gene family were split based on predicted secondary structure. Branch lengths of gene trees were reevaluated using generated alignment partitions, and insertion/deletion rates were recalculated for every branch of every gene phylogeny for each type of secondary structure (loops, alpha-helices, and beta-strands). As expected, most gene tree lineages show no insertion or deletion events. However, in full-length proteins and in loops, there is a more pronounced local maximum of density at about five insertions/deletions per unit of sequence divergence per unit of alignment length (fig. 3). Intuitively, longer insertions or deletions should have more pronounced effects on protein function than shorter ones. Therefore, we also analyzed events of different length in four groups: all indels and indels of minimal length 2, 3, and 4. Thus, 16 distributions of observed insertion/deletion rates were generated (fig. 3, supplementary fig. 4, Supplementary Material online).

Generation of Expected Distribution of Insertion/Deletion Rates

In order to be able to differentiate between varying strengths of selective pressure on indel substitutions, we need to have a null model of the process (occurrence of insertions and deletions under neutral selection) and underlying null distribution of insertion/deletion rates for every data set. To generate respective null distributions for every observed distribution, all observed insertion and deletion events were randomly reassigned to the branches of the gene phylogenies in proportion to the lengths of the branch and the alignment. Insertion/deletion rates were subsequently calculated for every branch based on the assigned number of events, branch length, and alignment length. The simulations were performed for the full-length protein data set and for the data set partitioned based on secondary structure, as well as for events of different length. Thus, 16 distributions of expected insertion/deletion rates were generated (fig. 3, supplementary fig. 4, Supplementary Material online). The percentile of every theoretical distribution corresponding to 50% of FDR was used as a threshold value to identify branches of gene phylogenies with significantly elevated insertion/deletion rates (supplementary table 1, Supplementary Material online). Our results showed that specific branches of many gene trees possess significantly higher number of insertions/deletions than would be expected by chance, further supporting the idea of natural selection acting on indel substitutions. For many partitions, the maximum observed event rate is several orders of magnitude higher than the maximum rate in randomized data. An insertion/deletion rate value significantly higher than that expected by chance on a particular branch of the gene phylogeny was interpreted as a sign of positive selection for insertion/deletion substitutions on that particular branch. The magnitude of the indel influence on the overall evolutionary trend might be estimated as a percentage of the branches where it was possible to detect positive selection on insertions or deletions (table 1). Insertions and deletions on up to 12% of all the branches in the data set evolved under positive selection.

Another outcome of our analysis was a set of branches with smaller number of indels than expected by chance; those branches would be assumed to carry indels under purifying selection. However, the lowest number of events that can be observed is zero (also resulting in a zero event rate). We took this value as the lowest possible threshold for detection of purifying selection. We almost always obtained a slightly larger number of observed branches with no indels than we would expect by chance. However, we could not identify the threshold corresponding to 50% FDR, as zero is the lowest threshold we can choose and it corresponds to at least 50th percentile of the expected distribution (supplementary table 2, Supplementary Material online).

The population biology of an organism should affect the extent to which random processes influence its evolution. Organisms of the phylum *Chlamydiae* are intracellular pathogens, thus they have smaller population sizes compared with free-living organisms found elsewhere in the

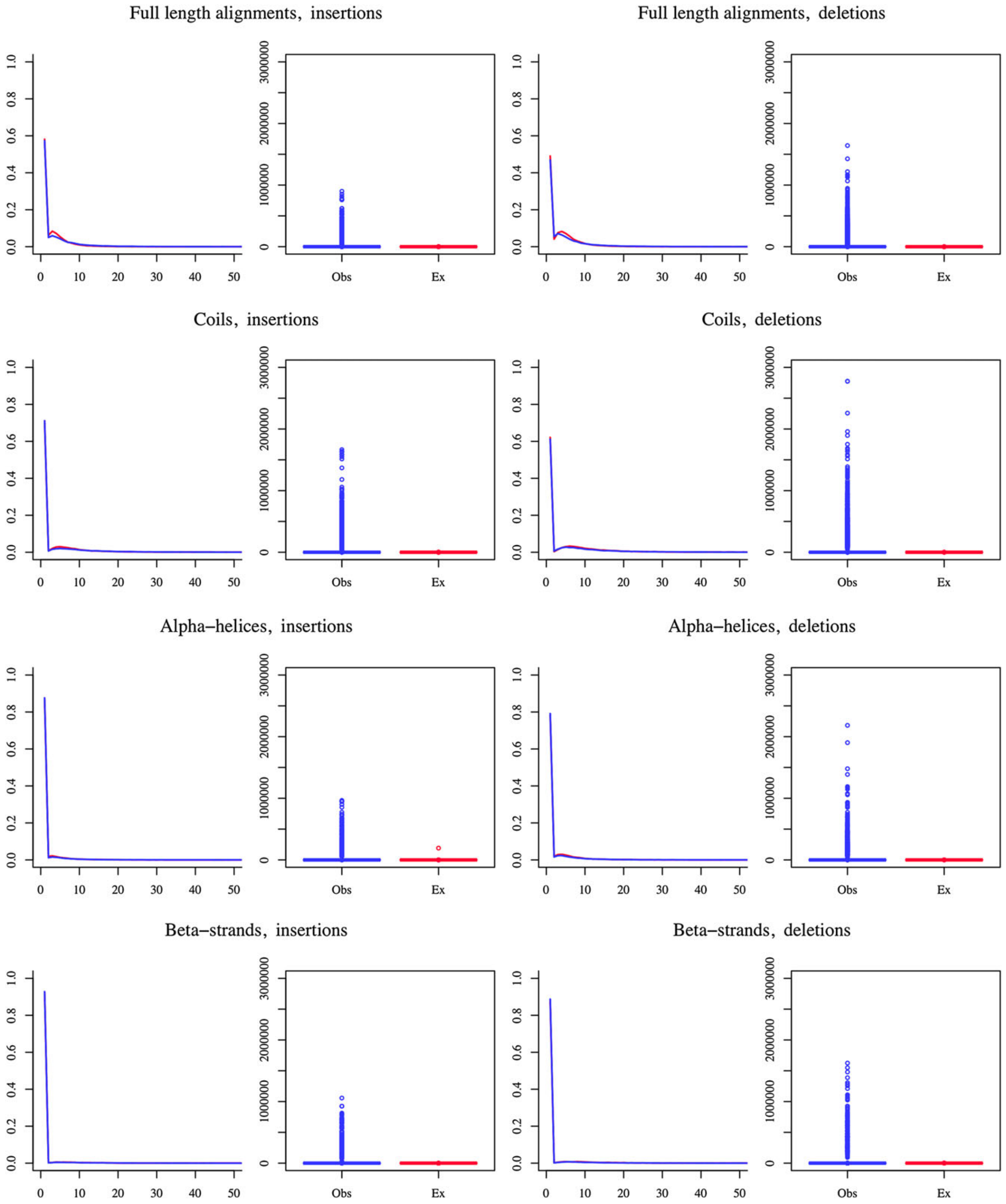


Fig. 3.—Insertion/deletion rate distributions, all events. Rate distributions are shown for every type of secondary structure (coils, alpha-helices, and beta-strands) and for full-length alignments. For every type of event (insertions and deletions), distributions are depicted with a histogram (*x* axes: event rates, number of events from 0 to 50 per unit of evolutionary distance, per unit of alignment length; *y* axes: density) and a boxplot of entire data set (*x* axes: class of the data; *y* axes: event rates, number of events per unit of evolutionary distance, per unit of alignment length). In both cases, blue and red colors denote observed and expected distributions, respectively.

Table 1

Number and percentage of branches showing evidence for positive selection on insertions and deletions in different length groups and secondary structural units

Full-length				
Total number of branches	52,018			
Length of accounted events	≥1	≥2	≥3	≥4
Number of branches with insertions under positive selection	6,466	4,858	3,059	2,018
Percentage of branches with insertions under positive selection	12.43	9.34	5.88	3.88
Number of branches with deletions under positive selection	6,683	5,130	3,607	2,455
Percentage of branches with deletions under positive selection	12.85	9.86	6.93	4.72
Coils				
Total number of branches	47,875			
Length of accounted events	≥1	≥2	≥3	≥4
Number of branches with insertions under positive selection	4,484	1,936	1,166	611
Percentage of branches with insertions under positive selection	9.37	4.04	2.44	1.28
Number of branches with deletions under positive selection	4,802	2,740	1,631	1,216
Percentage of branches with deletions under positive selection	10.03	5.72	3.41	2.54
α-helices				
Total number of branches	47,896			
Length of accounted events	≥1	≥2	≥3	≥4
Number of branches with insertions under positive selection	1,412	611	316	197
Percentage of branches with insertions under positive selection	2.95	1.28	0.66	0.41
Number of branches with deletions under positive selection	2,341	1,386	929	675
Percentage of branches with deletions under positive selection	4.89	2.89	1.94	1.41
β-strands				
Total number of branches	31,962			
Length of accounted events	≥1	≥2	≥3	≥4
Number of branches with insertions under positive selection	473	152	78	56
Percentage of branches with insertions under positive selection	1.48	0.48	0.24	0.18
Number of branches with deletions under positive selection	754	488	277	216
Percentage of branches with deletions under positive selection	2.36	1.53	0.87	0.68

superphylum. Therefore, we should observe a stronger effect of selection on the evolution of these free-living bacteria compared with chlamydiae. In order to evaluate the magnitude of selective effects on indels, we calculated the percentage of branches in which positive selection could be detected, performed both on branches leading to extant sequences from members of the phylum *Chlamydiae* and those leading to extant sequences from free-living organisms. The percentage of branches on which we were able to detect positive selection on insertion/deletion substitutions was always larger for free-living organisms than for chlamydiae, independent of the type of secondary structure under consideration or the minimal size of the substitution (supplementary table 3, Supplementary Material online). Thus, we confirmed our prediction of the effect of population biology on selection of indels.

In order to test the accuracy of the method in relation to the age of branches under consideration, we looked at the distribution of insertion/deletion rates on root-adjacent and tip-adjacent branches. As expected, we found more extreme insertion/deletion rate values on root-adjacent branches, even after resampling of the branches to achieve similar distributions of evolutionary distances (supplementary fig. 5, Supplementary Material online).

Evaluation of Algorithm Performance

Evaluation of the performance of new computational approaches is a critical step in the development of methodology. We lacked real data to evaluate the performance of our newly developed algorithm, thus we conducted a series of simulations to assess sensitivity and specificity of the approach. To simulate the data, 10% of the branches in the data set were randomly selected to be foreground branches (branches with indel substitutions under positive selection). Every branch was assigned to have an indel rate proportional to branch length, alignment length, and scaling factor sampled from a gamma distribution with different combination of parameters, as described in the Materials and Methods section. We used four different distributions to sample the scaling factor for foreground branches, in order to estimate the differences in the selective constraints needed to recognize the branch as carrying indels under positive selection. In the next step 100,000, 10,000, or 1,000 events were randomly assigned to different branches in the data set based upon the probabilities described above. Use of different numbers of events was relevant, as we observed different numbers of indels depending on the type of secondary structure under consideration and on the lengths of accounted events.

Table 2

FP and FN rates identified with a confidence level corresponding to 50% FDR, estimated for 12 data sets simulated under four different evolutionary scenarios with three different numbers of assigned events

Rate acceleration mode (*m*)		**1**		**2**		**3**		**4**	
Number of events		FP	FN	FP	FN	FP	FN	FP	FN
100,000	%	94.12	99.58	5.14	22.32	5.87	22.32	5.43	8.42
	Confidence value	0.00035		0.00409		0.00412		0.00483	
10,000	%	NA	NA	8.84	43.58	7.06	47.37	5.74	41.26
	Confidence value	NA		0.00309		0.00282		0.00311	
1,000	%	NA	NA	3.92	79.37	6.94	85.89	1.35	84.63
	Confidence value	NA		0.00107		0.00075		0.00077	

Note.—NA, not applicable.

We treated the resulting data sets as observed data and obtained distributions of event rates for every simulated data set. Using the randomization described above, we generated respective null distributions of event rates, determined threshold values corresponding to 50% FDR, and employed them to identify branches in the simulated data sets harboring significantly higher number of events compared with the null model. A priori knowledge of which branches carried indels under positive selection allowed identification of the corresponding FP and FN rates (table 2). It is intuitive that the larger the difference between the theoretical distributions underlying the partitions of simulated event rates, the more powerful should be the detection of positive selection. The best performance was observed on the data set with maximum separation of underlying theoretical distributions of event rates between the background and foreground branches ($m = 4$), and the largest number of assigned events (100,000) with the least stochasticity. The inferred FP rate was 5.43%, whereas the rate of FNs was 8.42%. In the cases of smaller number of events or smaller differences in rate between positive selection and neutral evolution, the FN rate became noticeably higher, whereas the rate of FPs remained approximately constant.

In order to evaluate the best possible performance of the algorithm, we estimated values of sensitivity and specificity for threshold values ranging from –0.01 ($0 - \delta$) to maximum rate value and derived corresponding ROC curves (supplementary fig. 6, Supplementary Material online). In the case of 100,000 assigned events, the performance of the algorithm depends on the number of observed events. It is possible to differentiate about 90% of all branches with an elevated rate of events (90% sensitivity) while having a very low level of FPs (more than 95% specificity). Performance goes down to about 75% and 40% sensitivity in the case of 10,000 and 1,000 assigned events, whereas still maintaining more than 95% specificity. Together, this control suggests that the approach is conservative. Even with an FDR of 50%, one should not expect many FPs but should expect to have large number of FNs especially with the lower number of observed events in α-helices and β-strands.

Applicability of Other Methods

Standard methods of evolutionary analysis aiming to detect positive Darwinian selection on amino acid replacement substitutions rely on the estimation of nonsynonymous to synonymous substitution rates ratio along branches of a gene phylogeny. These methods are sensitive to synonymous sites saturation if the sequences under consideration are too divergent. It is known that the upper limit of reliable K_s estimation is about 2 when estimated using PAML (Yang 2007). In order to assess the applicability of K_a/K_s-based approaches, we performed estimation of pairwise K_s values for all sequences in all gene families. We found that the average K_s value reached 22.8 synonymous substitutions per synonymous site (supplementary fig. 7, Supplementary Material online), which is far above the appropriate limit. Although some pairs of sequences in some gene families had acceptable K_s values, those sequences mostly originated from the same genome rather than from different genomes, and thus could not provide us with information on species divergence.

The most recently published methods of evolutionary rate-shift analysis (testing for type I functional divergence) enable large-scale analysis. However, RASER requires large data sets to reach appropriate statistical power. We tested whether RASER could detect a signal for evolutionary rate-shift in our data. We estimated log-likelihood for both the rate-shift enabling model and the model with no rate-shift, for every gene family. We used the likelihood ratio test to compare the two models under consideration. We considered chi-square distribution with degree of freedom equals three as a test statistics distribution, as suggested by the authors of RASER (Penn et al. 2008). However, we did not find any gene families that supported the rate-shift enabling model at an appropriate significance level. For most cases, the rate-shift enabling evolutionary model was statistically undistinguishable from the null model. From these results, we conclude that our newly developed method is the only method that allows characterization of indel evolution on a whole-genome scale as well as at large evolutionary distances.

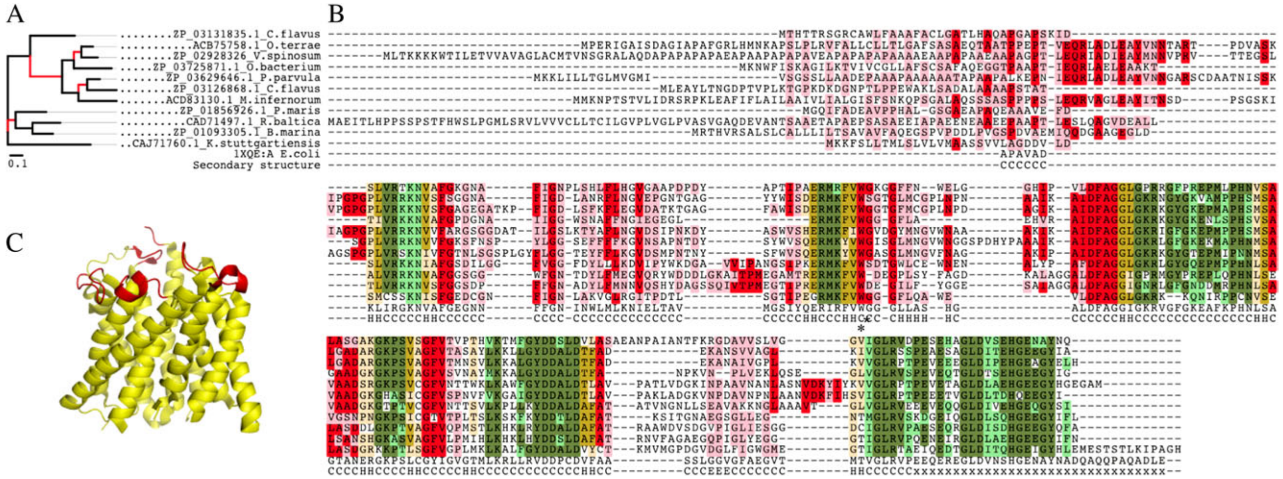

A
B
C
........ZP_03131835.1_C.flavus
..........ACB75758.1_O.terrae
........ZP_02928326_V.spinosum
.....ZP_03725871.1_O.bacterium
.......ZP_03629646.1_P.parvula
........ZP_03126868.1_C.flavus
.......ACD83130.1_M.infernorum
.........ZP_01856926.1_P.maris
..........CAD71497.1_R.baltica
........ZP_01093305.1_B.marina
..CAJ71760.1_K.stuttgartiensis
1XQE:A E.coli
Secondary structure
0.1
*

Fig. 4.—Ammonium transporter protein family. Gene family with insertions in coils under positive selection. (*A*) Phylogenetic tree of the ammonium transporter protein family (ion-coupled transporter, according to KEGG), individual sequences are designated by GenBank accession numbers and species name. Branches with significantly high level of insertions in coils are shown in red. Two additional names correspond to PDB accession numbers for the homologous sequence with determined tertiary structure and to the corresponding secondary structural elements identified based on tertiary structure. (*B*) Corresponding multiple sequence alignment of the members of the protein family and *Escherichia coli* AmtB sequence from PDB (the parts of alignment corresponding to transmembrane helices have been trimmed). Bright and light shadings correspond to 50% identical or similar (based on PAM250) residues in the sequences of protein family. Red palette, periplasmic parts; yellow palette, trimmed transmembrane parts; green palette, cytoplasmic parts. Last line represents types of secondary structure elements, based on tertiary structure of *E. coli* AmtB 1XQE:A (Zheng et al. 2004). Residue W148 at the beginning of periplasmic coil between transmembrane helices four and five is marked with *. (*C*) Structural model 1XQE:A of *E. coli* AmtB ammonium transporter was used to show periplasmic side coiled and α-helical regions with unexpectedly high number of insertions (marked in red).

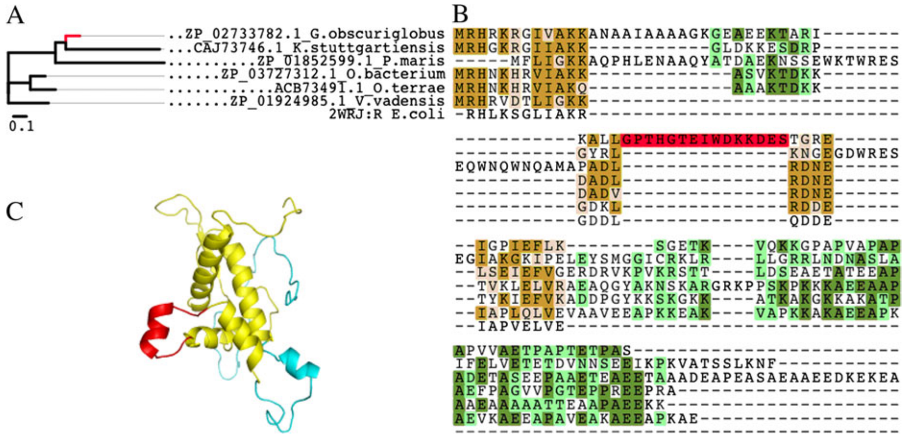
Frequency and Length Distributions of Insertions and Deletions

This study represents the first analysis of indel substitutions in the genomes of distantly related organisms, providing insights into the general characteristics of insertions and deletions in the set of divergent protein sequences, as well as into their patterns of selective constraints. Using the workflow described above, we identified 37,365 insertion and 53,557 deletion events along the branches of the gene trees in full-length alignments. Observing larger number of deletions than insertions is consistent with what has been shown in other studies of protein-coding sequences from nematodes (Wang et al. 2009) and in a rat/mouse comparison (Taylor et al. 2004). It seems that the presence of small genomes from chlamydial species might have influenced our results for insertion/deletion frequency; it has been shown in eukaryotes that DNA loss is one of the underlying mechanisms of genome shrinkage (Petrov 2002). However, here, we examine the evolution of individual genes, whereas processes associated with dramatic genome size changes in pathogenic bacteria occur on a larger scale with loss of whole genes or large parts of genomes containing several open reading frames (Mira et al. 2001; Moran and Mira 2001; Gregory 2004; Nilsson et al. 2005).

The observed length distributions of insertions and deletions in different types of secondary structure are shown in supplementary figure 8 (Supplementary Material online). The longest insertion identified in our data set was 217 amino acids, whereas the longest deletion was 190 amino acids. The most common insertion or deletion event was a one amino acid-long substitution, independent of the type of secondary structure under consideration. The mean length value of observed insertions/deletions was 3.77/3.22 amino acids for full-length proteins. Observed insertions generally tended to be longer than deletions in all the types of structural elements.

Mapping of Indels on Cellular Pathways

Selection is observed at the level of the individual gene/protein but actually occurs in the context of broader cellular biology. We used KEGG (Kanehisa et al. 2010) metabolic pathways to classify gene families in the data set and systematically identify molecular pathways affected by indel processes. We linked every gene family with information from the KEGG molecular pathways database using a Blast search against the database. We were able to map all full-length gene families onto 106 groups of cellular pathways. However, the total number of pathways obtained varied depending on the specific types of secondary structure in which indels occurred (supplementary fig. 9, Supplementary Material online). We employed a binomial test to identify pathways consistently overrepresented among


A
..ZP_02733782.1_G.obscuriglobus
...CAJ73746.1_K.stuttgartiensis
..........ZP_01852599.1_P.maris
......ZP_03727312.1_O.bacterium
............ACB73491.1_O.terrae
.......ZP_01924985.1_V.vadensis
2WRJ:R E.coli
0.1
B
MRHRKRGIVAKKANAAIAAAAGKGEAEEKTARI-------
MRHGKRGIIAKK------------GLDKKESDRP--------
-----MFLIGKKAQPHLENAAQYATDAEKNSSEWKTWRES
MRHNKHRVIAKK--------------ASVKTDKK-------
MRHNKHRVIAKQ--------------AAAKTDKK-------
MRHRVDTLIGKK----------------------------
-RHLKSGLIAKR----------------------------
------------KALLGPTHGTEIWDKKDESTGRE-------
------------GYRL---------------KNGEGDWRES
EQWNQWNQAMAPADL---------------RDNE-------
------------DADL---------------RDNE-------
------------DADV---------------RDNE-------
------------GDKL---------------RDDE-------
------------GDDL---------------QDDE-------
--IGPIEFLK---------SGETK----VQKKGPAPVAPAP
EGIAKGKIPELEYSMGGICRKLR----LLGRRLNDNASLA
--LSEIEFVGERDRVKPVKRSTT----LDSEAETATEEAP
--TVKLELVRAEAQGYAKNSKARGRKPPSKPKKKAEEAAP
--TYKIEFVKADDPGYKKSKGKK----ATKAKGKKAKATP
--IAPLQLVEVAAVEEAPKKEAK----VAPKKAKAEEAPK
--IAPVELVE------------------------------
APVVAETPAPTETPAS-------------------------
IFELVETETDVNNSEEIKPKVATSSLKNF-----------
ADETASEEPAAETEAEETAAADEAPEASAEAAEEDKEKEA
AEFPAGVVPGTEPPREEPRA--------------------
AAEAAAAATTEAAPAEEKK---------------------
AEVKAEEAPAVEAKAEEAPKAE------------------

C

Fig. 5.—Ribosomal protein L17 gene family. Gene family with insertions in alpha-helices under positive selection. Phylogenetic tree of ribosomal protein L17 gene family (ribosomes, according to KEGG), individual sequences are designated by GenBank accession numbers and species name. Branches with significantly high level of insertions are shown in red. Additional name corresponds to PDB accession number for the homologous sequence with determined tertiary structure. (*A*) Corresponding multiple sequence alignment of the members of protein family and *Escherichia coli* ribosomal protein L17 sequence from PDB (the parts of alignment corresponding to common protein segments have been trimmed). Bright and light shadings correspond to 50% identical or similar (based on PAM250) residues in the sequences of protein family. Yellow palette, trimmed parts shared with *E. coli* ribosomal protein L17; green palette, parts shared by more than one member of the family. Red shading, *Gemmata obscuriglobus* specific insertion. (*A*) Structural model of *G. obscuriglobus* ribosomal protein L17, constructed using i-TASSER fold recognition server based on 2WRJ:R tertiary structure (Gao et al. 2009). Common core is shown in yellow; parts shared by more than one member of the family are shown in cyan; *G. obscuriglobus*-specific insertion is shown in red.

gene families with positive selection on insertions/deletions of different length in varying secondary structural elements. Different types of transporters (ABC transporters, pore ion channels) as well as several pathways related to general metabolism (cysteine and methionine, thiamine, selenoamino acid, phenylalanine, sphingolipid metabolism, base excision repair, glycosaminoglycan degradation, terpenoid backbone biosynthesis, ribosome, bacterial secretion systems, and protein export) were consistently overrepresented among gene families with positive selection on insertions/deletions of different length (supplementary fig. 9*A*, Supplementary Material online). Noticeably, ABC type transporters and ion-coupled transporters show elevated rates of deletions and insertions in coils. This may suggest a general pattern of evolution for these types of proteins. Insertions (deletions) in coiled regions might change the structural composition of the protein by introducing (eliminating) structural elements in the case of long indels containing alpha-helices or beta-strands. In the case of indels that do not affect structural composition of the protein, they may alter flexibility of the existing protein fold in terms of positioning of structural elements relative to each other or to binding partners. In some cases, this might also change the thermodynamic stability of proper protein folding (Viguera and Serrano 1997; Meenan et al. 2010). As described below, we examined the structural and functional consequences of indel events in example gene families that exhibited evidence for positive selection on indel substitutions.

One of the ion-coupled transporters with an unexpectedly high number of insertions in loop regions is the ammonium transporter from planctomycete and verrucomicrobia species (fig. 4). Several branches of the gene phylogeny for this protein family exhibit elevated levels of insertions in coils. Additionally, mapping of insertions on the tertiary structure of *E. coli* AmtB showed clustering of otherwise conserved insertions in periplasmic loops. There are no known binding partners that would interact with the periplasmic domain of AmtB. However, a previous study of the *E. coli* protein allowed identification of several mutations in the periplasmic domain of the pore entrance that significantly increased ammonium uptake (Javelle et al. 2008). W148A is particularly interesting as it is located in the periplasmic coil between the fourth and fifth transmembrane helices, adjacent to a small periplasmic helical element. In the proteins of the planctomycete and verrucomicrobia clade, the periplasmic helix contained several small indels. Furthermore, part of the loop adjacent to the fifth transmembrane helix contained an additional protein segment conserved among members of the family. Considering that proteins in this gene family originate from organisms living in low-nutrient environments, it is possible that the observed insertions would facilitate evolution of a more efficient ammonium transporter. Although no particular molecular function of the identified inserts has been proven at this point, we hypothesize that the observed insertions might underlie the emergence of a new regulatory interaction on the periplasmic side or might be associated with altering the efficiency of transport.

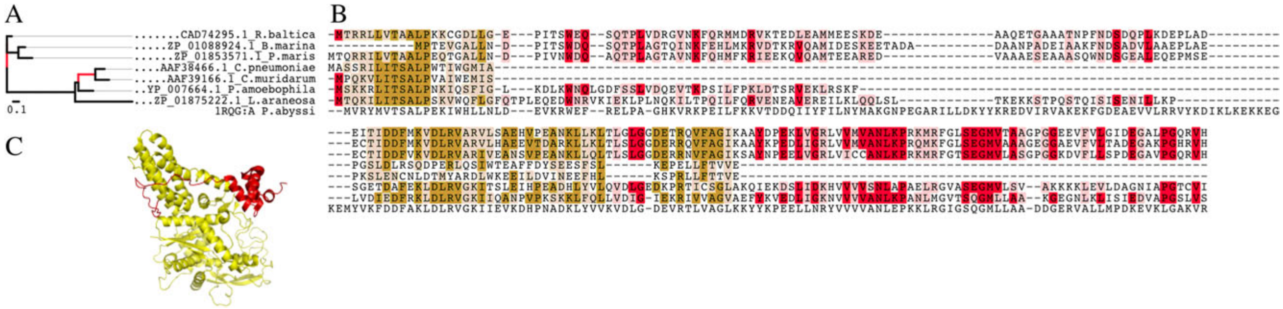

A
.......CAD74295.1_R.baltica
.....ZP_01088924.1_B.marina
......ZP_01853571.1_P.maris
....AAF38466.1_C.pneumoniae
.....AAF39166.1_C.muridarum
..YP_007664.1_P.amoebophila
...ZP_01875222.1_L.araneosa
1RQG:A P.abyssi
0.1
B
-MTRRLLVTAALPKKCGDLLG-E---PITSWEQ--SQTPLVDRGVNKFQRMMDRVKTEDLEAMMEESKDE--------------AAQETGAAATNPFNDSDQPLKDEPLAD---------
-------------MPTEVGALLN-D---PITSWDQ--SQTPLAGTQINKFEHLMKRVDTKRVQAMIDESKEETADA----------DAANPADEIAAKFNDSADVLAAEPLAE---------
MTQRRILVTAALPEQTGALLN-D---PIVNWDQ--AQTPLAGTAVNKFQHMFKRIEEKQVQAMTEEARED--------------VAAAESEAAASQWNDSGEALEQEPMSE---------
MASSRILITSALPWTIWGMIA---
-MPQKVLITSALPVAIWEMIS---
-MSKKRLITSALPNKIQSFIG-L---KDLKWNQLGDFSSLVDQEVTKPSILFPKLDTSRVEKLRSKF--
-MTQKILITSALPSKVWQFLGFQTPLEQEDWNRVKIEKLPLNQKILTPQILFQRVENEAVEREILKLQQLSL-----------TKEKKSTPQSTQISISENILLKP-------------
--MVRYMVTSALPEKIWHLLNLD---EVKRWEF--RELPA-GHKVRKPEILFKKVTDDQIIYFILNYMAKGNPEGARILLDKYYKREDVIRVAKEKFGDEAEVVLRRVYKDIKLKEKKEG
---EITIDDFMKVDLRVARVLSAEHVPEANKLLKLTLGLGGDETRQVFAGIKAAYDPEKLVGRLVVMVANLKPRKMRFGLSEGMVTAAGPGGEEVFVLGIDEGALPGQRVH
---ECTIDDFMKVDLRVARVLHAEEVTDARKLLKLTLSLGGDERRQVFAGIKAAYKPEDLIGRLVVMVANLKPRQMKFGLSEGMVAAAGEGGAEVFVLTADEGAKPGHRVH
---ECTIDDFVKVDLRVARIVEANSVPEANKLLQLTLSLGGDERRNVFAGIKSAYNPEELVGRLVICCANLKPRKMRFGTSEGMVLASGPGGKDVFLLSPDEGAVPGQRVH
---PGSLDLRSQDPERLQSIWTEAFFDYSEESFSL-------KEPELLFTVVE--
---PKSLENCNLDTMYARDLWKEEILDVINEEFHL-------KSPRLLFTTVE--
---SGETDAFEKLDLRVGKITSLEIHPEADHLYVLQVDLGEDKPRTICSGLAKQIEKDSLIDKHVVVVSNLAPAELRGVASEGMVLSV--AKKKKLEVLDAGNIAPGTCVI
---LVDIEDFRKLDLRVGKIIQANPVPKSKKLFQLLVDIG-IEKRIVVAGVAEFYKVEDLIGKNVVVVANLKPANLMGVTSQGMLLAA--KGEGNLKLISIEDVAPGSLVS
KEMYVKFDDFAKLDLRVGKIIEVKDHPNADKLYVVKVDLG-DEVRTLVAGLKKYYKPEELLNRYVVVVANLEPKKLRGIGSQGMLLAA-DDGERVALLMPDKEVKLGAKVR
C

Fig. 6.—Methionyl-tRNA synthetase protein family. Gene family with deletions in full-length proteins under positive selection. (*A*) Phylogenetic tree of methionyl-tRNA synthetase gene family (selenoamino acid metabolism, according to KEGG), individual sequences are designated by GenBank accession numbers and species name. Branches with significantly high level of deletions are shown in red. Additional name corresponds to PDB accession number for the homologous sequence with determined tertiary structure. (*A*) Corresponding multiple sequence alignment of the members of the protein family and *Pyrococcus abyssi* methionyl-tRNA synthetase sequence from PDB (the parts of alignment corresponding to common protein segments have been trimmed). Bright and light shadings correspond to 50% identical or similar (based on PAM250) residues in the sequences of protein family. Yellow palette, trimmed parts shared with *P. abyssi* methionyl-tRNA synthetase; red palette, otherwise conserved parts which were deleted from *Chlamydia muridarum* and *Chlamydophila pneumoniae* proteins. (*A*) Structural model of 1RQG:A *P. abyssi* methionyl-tRNA synthetase (Crepin et al. 2004). Common core is shown in yellow; otherwise conserved regions that have undergone deletions in *C. muridarum* and *C. pneumoniae* are shown in red.

Gemmata obscuriglobus Evolutionary Lineage

Planctomycetes and verrucomicrobia share a feature that is unusual for bacteria: a compartmentalized cell plan with at least one additional intracellular membrane, which separates ribosome-containing riboplasm from ribosome-free paryphoplasm (Fuerst 2005; Lee et al. 2009). *Gemmata obscuriglobus* cells have an especially unusual cellular organization featuring an extra compartment surrounded by a double-layered membrane envelope and containing condensed genomic DNA (Fuerst and Webb 1991). We identified gene families where we observed an unexpectedly high number of indels on the *G. obscuriglobus* lineage, and we hypothesize that these may be candidate genes supporting the unique biology that emerged on this lineage. Several metabolic pathways were overrepresented among these gene families (supplementary fig. 9*B*, Supplementary Material online). Some pathways were consistently overrepresented in this data set, among these we found different types of general metabolic pathways (cysteine and methionine, sulfur, lipoic acid metabolism, protein export, ATPases, oxidative phosphorylation, fatty acid, peptidoglycan, pantothenate and CoA biosynthesis, pores ion channels, bacterial secretion systems, ribosome). Some of these families have a plausible connection to the unusual cell structures. For instance, it is known that planctomycetes and chlamydiae do not contain peptidoglycan in their cell wall (Pilhofer et al. 2008). It is possible that in *G. obscuriglobus*, genes normally involved in peptidoglycan biosynthesis have evolved a new molecular function and indels might have contributed to this process.

Additionally, one of the hypotheses concerning the specific physiology of *G. obscuriglobus* is that the extra compartment with enclosed genomic DNA is an analog of the eukaryotic nucleus, and the double-layered membrane envelope facilitates the uncoupling of transcription and translation (Fuerst 2005). Along these lines, indels in ribosomal proteins might have contributed to changes from the classical bacterial translation process. The L17 ribosomal protein provides an example of a ribosomal protein with an elevated insertion rate in alpha-helices on the *G. obscuriglobus* lineage. Further examination revealed large insertions in other members of the protein family compared with *E. coli* proteins (fig. 5). According to a tertiary structure prediction, both the *G. obscuriglobus* and planctomycete lineages show specific insertions in alpha-helical elements in close proximity to regions of the 23S rRNA (data not shown), whereas interactions with other ribosomal proteins seem to be unaffected. This finding is striking given that ribosomal proteins are considered to be among the most evolutionarily conserved.

Chlamydiae Evolutionary Lineage

Organisms of the phylum *Chlamydiae* are important pathogens causing sexually transmitted and pulmonary diseases in humans and other mammals as well as infecting lower eukaryotes (Horn 2008). Organisms of the phylum *Chlamydiae*, like many other intracellular pathogens, have undergone genome reduction compared with free-living bacteria. A series of changes in lifestyle occurred on different evolutionary lineages leading to and within the chlamydial clade. For instance, it would be expected based on parsimony that the lineage leading from the common ancestor of *Lentisphaerae* and *Chlamydiae* to the ancestor of all the *Chlamydiae* species would represent the transition from a free-living/commensal lifestyle to an obligate host association, as *Lentisphaerae* and the deeper-branching planctomycetes are free-living environmental organisms. A medically important lifestyle shift also occurred when ancient chlamydiae transitioned from a simple to a multicellular eukaryotic host (Horn 2008). In relation to the current species set, this transition took place on the lineage leading from the common ancestor of all chlamydiae to the ancestor of *C. pneumoniae* and *C. muridarum* (fig. 1). Determination of specific biological pathways affected by various evolutionary processes on lineages in the chlamydial clade will provide more detailed information on the acquisition of pathogenicity and the development of host specificity by chlamydiae. We identified several biological pathways, which were overrepresented (supplementary fig. 9C, Supplementary Material online) on branches of the chlamydial clade with unexpectedly high level of insertions and deletions. Unlike the entire data set, we did not observe transporter proteins to be significantly affected by indels on the lineages of chlamydial clade. Most of the consistently overrepresented categories are general biochemical pathways. One example of a gene family with observed high level of deletions on the lineage leading to the common ancestor of *C. pneumoniae* and *C. muridarum* is the methionyl-tRNA synthetase protein family (fig. 6). The C-terminal dimerization domain was deleted from *C. pneumoniae* and *C. muridarum* but retained in all the other proteins of the gene family. This implies that in all the other organisms represented in the gene family, including *P. amoebophila*, methionyl-tRNA synthetase acts as a homodimer, which is shown to have higher affinity to tRNA (Crepin et al. 2002).

Summary

Most published comparative functional genomic and molecular evolutionary analysis focuses on the dynamics and functional consequences of amino acid replacements in proteins. Although it is clear that insertions and deletions contribute to changes in protein function and evolve under the control of selection, genetic variations of this type have been much less well studied due to the lack of sensitive and reliable methods of analysis. Here, we developed an approach to study insertions and deletions in a genome-wide manner and revealed patterns of evolutionary constraints on insertion/deletion substitutions on various branches of individual gene phylogenies. In general, we observed more deletions than insertions, however, observed insertions tend to be longer than deletions. Indel substitutions preferentially occurred in coiled regions. We also found evidence for positive selection of indel substitutions on up to 12% of branches in the entire data set. Using simulations, we show that the approach developed here is conservative and should not yield a significant number of FP results. Lastly, we have provided information on functional and structural variations in highly divergent sequences from a remarkable but evolutionarily understudied group of organisms in the PVC superphylum. Identified changes may be connected to the development of complex intracellular and extracellular structures observed in planctomycetes and verrucomicrobia and lifestyle shifts in chlamydiae.

Supplementary Material

Supplementary figures 1–9 and tables 1–3 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

The authors thank C. Alex Buerkle, Stormy Knight, Grigory Kolesov, and Jessica Siltberg-Liberles for helpful discussions. This work was supported by the National Institutes of Health (P20 RR016474 to O.K.K.] and National Science Foundation (NSF) (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. No unpublished data or data obtained from personal communications were used in the present study. The genomes analyzed in this study are already deposited in GenBank. Original and modified gene family alignments are available on request.

Literature Cited

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

Abhiman S, Sonnhammer ELL. 2005. Large-scale prediction of function shift in protein families with a focus on enzymatic function. Proteins. 60:758–768.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol. 215:403–410.

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

Benner SA, Gerloff D. 1991. Patterns of divergence in homologous proteins as indicators of secondary and tertiary structure: a prediction of the structure of the catalytic domain of protein kinases. Adv Enzyme Regul. 31:121–181.

Benner SA, Trabesinger N, Schreiber D. 1998. Post-genomic science: converting primary structure into physiological function. Adv Enzyme Regul. 38:155–180.

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

Blouin C, Boucher Y, Roger AJ. 2003. Inferring functional constraints and divergence in protein families using 3D mapping of phylogenetic information. Nucleic Acids Res. 31:790–797.

Brandstrom M, Ellegren H. 2007. The genomic landscape of short insertion and deletion polymorphisms in the chicken (Gallus gallus) genome: a high frequency of deletions in tandem duplicates. Genetics. 176:1691–1701.

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

Britten RJ, Rowen L, Williams J, Cameron RA. 2003. Majority of divergence between closely related DNA samples is due to indels. Proc Natl Acad Sci U S A. 100:4661–4665.

Chan S, Hsing M, Hormozdiari F, Cherkasov A. 2007. Relationship between insertion/deletion (indel) frequency of proteins and essentiality. BMC Bioinformatics. 8:227.

Chang MSS, Benner SA. 2004. Empirical analysis of protein insertions and deletions determining parameters for the correct placement of gaps in protein sequence alignments. J Mol Biol. 341:617–631.

Chen CH, Chuang TJ, Liao BY, Chen FC. 2010. Scanning for the signatures of positive selection for human-specific insertions and deletions. Genome Biol Evol. 2009:415–419.

Cho J, Vergin KL, Morris RM, Giovannoni SJ. 2004. Lentisphaera araneosa gen. nov., sp. nov, a transparent exopolymer producing marine bacterium, and the description of a novel bacterial phylum, Lentisphaerae. Environ Microbiol. 6:611–621.

Clark AG, et al. 2003. Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios. Science. 302:1960–1963.

Crepin T, Schmitt E, Blanquet S, Mechulam Y. 2002. Structure and function of the C-terminal domain of methionyl-tRNA synthetase. Biochemistry. 41:13003–13011.

Crepin T, Schmitt E, Blanquet S, Mechulam Y. 2004. Three-dimensional structure of methionyl-tRNA synthetase from Pyrococcus abyssi. Biochemistry. 43:2635–2644.

Davids W, Gamieldien J, Liberles DA, Hide W. 2002. Positive selection scanning reveals decoupling of enzymatic activities of carbamoyl phosphate synthetase in Helicobacter pylori. J Mol Evol. 54:458–464.

Dayhoff MO, Schwartz RM, Orcutt BC. 1978. A model of evolutionary change in proteins. Atlas of protein sequence and structure. Washington (DC): National Biomedical Research Foundation. 5:345–352.

DeLano WL. 2002. The PyMOL user's manual. San Carlos (CA): DeLano Scientific.

Dorman KS. 2007. Identifying dramatic selection shifts in phylogenetic trees. BMC Evol Biol. 7(1 Suppl): S10.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32:1792–1797.

Edwards RJ, Shields DC. 2004. GASP: gapped ancestral sequence prediction for proteins. BMC Bioinformatics. 5:123.

Embley MT, Hirt RP, Williams DM. 1994. Biodiversity at the molecular level: the domains, kingdoms and phyla of life. Philos Trans R Soc Lond B Biol Sci. 345:21–33.

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.

Fitch WM. 1971. Toward defining the course of evolution: minimum change for a specific tree topology. Syst Zool. 20:406–416.

Fuerst JA. 2005. Intracellular compartmentation in planctomycetes. Annu Rev Microbiol. 59:299–328.

Fuerst JA, Webb RI. 1991. Membrane-bounded nucleoid in the eubacterium Gemmatata obscuriglobus. Proc Natl Acad Sci U S A. 88:8184–8188.

Gao YG, et al. 2009. The structure of the ribosome with elongation factor G trapped in the posttranslocational state. Science. 326:694–699.

Gregory TR. 2004. Insertion-deletion biases and the evolution of genome size. Gene. 324:15–34.

Griffiths E, Gupta RS. 2007. Phylogeny and shared conserved inserts in proteins provide evidence that Verrucomicrobia are the closest known free-living relatives of chlamydiae. Microbiology. 153:2648–2654.

Gu X. 1999. Statistical methods for testing functional divergence after gene duplication. Mol Biol Evol. 16:1664–1674.

Gu X. 2001. Maximum-likelihood approach for gene family evolution under functional divergence. Mol Biol Evol. 18:453–464.

Gu X, Fu YX, Li WH. 1995. Maximum likelihood estimation of the heterogeneity of substitution rate among nucleotide sites. Mol Biol Evol. 12:546–557.

Hedlund BP, Gosink JJ, Staley JT. 1997. Verrucomicrobia div. nov., a new division of the bacteria containing three new species of Prosthecobacter. Antonie van Leeuwenhoek. 72:29–38.

Henikoff S, Henikoff JG. 1992. Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci U S A. 89:10915–10919.

Horn M. 2008. Chlamydiae as symbionts in eukaryotes. Annu Rev Microbiol. 62:113–131.

Huson DH, Bryant D. 2006. Application of phylogenetic networks in evolutionary studies. Mol Biol Evol. 23:254–267.

Javelle A, et al. 2008. Substrate binding, deprotonation, and selectivity at the periplasmic entrance of the Escherichia coli ammonia channel AmtB. Proc Natl Acad Sci U S A. 105:5040–5045.

Jenkins C, Fuerst JA. 2001. Phylogenetic analysis of evolutionary relationships of the planctomycete division of the domain bacteria based on amino acid sequences of elongation factor Tu. J Mol Evol. 52:405–418.

Jones DT. 1999. Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol. 292:195–202.

Jones DT, Taylor WR, Thornton JM. 1992. The rapid generation of mutation data matrices from protein sequences. Comput Appl Biosci. 8:275–282.

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.

Knudsen B, Miyamoto MM. 2001. A likelihood ratio test for evolutionary rate shifts and functional divergence among proteins. Proc Natl Acad Sci U S A. 98:14512–14517.

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 MJ. 2007. Evolution of the core and pan-genome of Streptococcus: positive selection, recombination, and genome composition. Genome Biol. 8(5): R71.

Lefebure T, Stanhope MJ. 2009. Pervasive, genome-wide positive selection leading to functional divergence in the bacterial genus Campylobacter. Genome Res. 19:1224–1232.

Li L, Stoeckert CJ, Roos DS. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. 13:2178–2189.

Liberles DA. 2001. Evaluation of methods for determination of a reconstructed history of gene sequence evolution. Mol Biol Evol. 18:2040–2047.

Liberles DA, Schreiber DR, Govindarajan S, Chamberlin SG, Benner SA. 2001. The adaptive evolution database (TAED). Genome Biol. 2:research0028.1–research0028.6.

Lindsay MR, et al. 2001. Cell compartmentalisation in planctomycetes: novel types of structural organisation for the bacterial cell. Arch Microbiol. 175:413–429.

Lonhienne TGA, et al. 2010. Endocytosis-like protein uptake in the bacterium Gemmata obscuriglobus. Proc Natl Acad Sci U S A. 107:12883–12888.

Meenan NAG, et al. 2010. The structural and energetic basis for high selectivity in a high-affinity protein–protein interaction. Proc Natl Acad Sci U S A. 107:10080–10085.

Messier W, Stewart CB. 1997. Episodic adaptive evolution of primate lysozymes. Nature. 385:151–154.

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

Moran NA, Mira A. 2001. The process of genome shrinkage in the obligate symbiont Buchnera aphidicola. Genome Biol. 2:research0054.1–research0054.12.

Nilsson AI, et al. 2005. Bacterial genome size reduction by experimental evolution. Proc Natl Acad Sci U S A. 102:12112–12116.

Orsi RH, Sun Q, Wiedmann M. 2008. Genome-wide analyses reveal lineage specific contributions of positive selection and recombination to the evolution of Listeria monocytogenes. BMC Evol Biol. 8:233.

Osterberg MK, Shavorskaya O, Lascoux M, Lagercrantz U. 2002. Naturally occurring indel variation in the Brassica nigra COL1 gene is associated with variation in flowering time. Genetics. 161:299–306.

Pascarella S, Argos P. 1992. Analysis of insertions/deletions in protein structures. J Mol Biol. 224:461–471.

Penn O, et al. 2008. Evolutionary modeling of rate shifts reveals specificity determinants in HIV-1 subtypes. PLoS Comput Biol. 4:e1000214.

Petersen L, Bollback JP, Dimmic M, Hubisz M, Nielsen R. 2007. Genes under positive selection in Escherichia coli. Genome Res. 17:1336–1343.

Petrov DA. 2002. DNA loss and evolution of genome size in Drosophila. Genetica. 115:81–91.

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.

Podlaha O, Webb DM, Tucker PK, Zhang J. 2005. Positive selection for indel substitutions in the rodent sperm protein Catsper1. Mol Biol Evol. 22:1845–1852.

Podlaha O, Zhang J. 2003. Positive selection on protein-length in the evolution of a primate sperm ion channel. Proc Natl Acad Sci U S A. 100:12241–12246.

Pupko T, Galtier N. 2002. A covarion-based method for detecting molecular adaptation: application to the evolution of primate mitochondrial genomes. Proc Biol Sci. 269:1313–1316.

Reeves JH. 1992. Heterogeneity in the substitution process of amino acid sites of proteins coded for by mitochondrial DNA. J Mol Evol. 35:17–31.

Retchless AC, Lawrence JG. 2010. Phylogenetic incongruence arising from fragmented speciation in enteric bacteria. Proc Natl Acad Sci U S A. 107:11453–11458.

Rivera MC, Jain R, Moore JE, Lake JA. 1998. Genomic evidence for two functionally distinct gene classes. Proc Natl Acad Sci U S A. 95:6239–6244.

Roenner S, Liesack W, Wolters J, Stackebrandt E. 1991. Cloning and sequencing of a large fragment of the atpD gene of Pirellula marine—a contribution to the phylogeny of Planctomycetales. Endocyto Cell Res. 7:219–229.

Roy A, Kucukural A, Zhang Y. 2010. I-TASSER: a unified platform for automated protein structure and function prediction. Nat Protoc. 5:725–738.

Santarella-Mellwig R, et al. 2010. The compartmentalized bacteria of the planctomycetes-verrucomicrobia-chlamydiae superphylum have membrane coat-like proteins. PLoS Biol. 8:e1000281.

Schloss PD, Handelsman J. 2004. Status of the microbial census. Microbiol Mol Biol Rev. 68:686–691.

Schully SD, Hellberg ME. 2006. Positive selection on nucleotide substitutions and indels in accessory gland proteins of the Drosophila pseudoobscura subgroup. J Mol Evol. 62:793–802.

Smith JM, Smith NH. 1996. Synonymous nucleotide divergence: what is "saturation"? Genetics. 142:1033–1036.

Sorek R, et al. 2007. Genome-wide experimental determination of barriers to horizontal gene transfer. Science. 318:1449–1452.

Stackebrandt E, et al. 1984. Molecular genetic evidence for early evolutionary origin of budding peptidoglycan-less eubacteria. Nature. 307:735–737.

Stamatakis A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics. 22:2688–2690.

Taylor MS, Ponting CP, Copley RR. 2004. Occurrence and consequences of coding sequence insertions and deletions in mammalian genomes. Genome Res. 14:555–566.

Van de Peer Y, Neefs JM, De Rijk P, De Vos P, De Wachter R. 1994. About the order of divergence of the major bacterial taxa during evolution. Syst Appl Microbiol. 17:32–38.

Viguera AR, Serrano L. 1997. Loop length, intramolecular diffusion and protein folding. Nat Struct Mol Biol. 4:939–946.

Wagner M, Horn M. 2006. The planctomycetes, verrucomicrobia, chlamydiae and sister phyla comprise a superphylum with biotechnological and medical relevance. Curr Opin Biotechnol. 17:241–249.

Wang Z, et al. 2009. Systematic analysis of insertions and deletions specific to nematode proteins and their proposed functional and evolutionary relevance. BMC Evol Biol. 9:23.

Ward NL, et al. 2000. Comparative phylogenetic analyses of members of the order Planctomycetales and the division Verrucomicrobia: 23S rRNA gene sequence analysis supports the 16S rRNA gene sequence-derived phylogeny. Int J Syst Evol Microbiol. 50:1965–1972.

Ward NL, Staley JT, Fuerst JA, Giovannoni S, Schlesner H, Stackebrandt E. 2006. The order *Planctomycetales*, including the genera *Planctomyces, Pirellula, Gemmata* and *Isosphaera* and the *Candidatus* genera *Brocadia, Kuenenia* and *Scalindua*. In: Dworkin M, Falkow S, Rosenberg E, Schleifer KH, Stackebrandt E, editors. The Prokaryotes: a Handbook on the Biology of Bacteria. New York: Springer. pp. 757–793.

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.

Yang Z. 1993. Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. Mol Biol Evol. 10:1396–1401.

Yang Z. 2006. Computational molecular evolution. Oxford: Oxford University Press.

Yang Z. 2007. PAML 4: Phylogenetic Analysis by Maximum Likelihood. Mol Biol Evol. 24:1586–1591.

Yang Z, Nielsen R. 2002. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. Mol Biol Evol. 19:908–917.

Zhang J, Nielsen R, Yang Z. 2005. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. Mol Biol Evol. 22:2472–2479.

Zheng L, Kostrewa D, Bernèche S, Winkler FK, Li XD. 2004. The mechanism of ammonia transport based on the crystal structure of AmtB of Escherichia coli. Proc Natl Acad Sci U S A. 101:17090–17095.

Zhou T, Gu W, Wilke CO. 2010. Detecting positive and purifying selection at synonymous sites in yeast and worm. Mol Biol Evol. 27:1912–1922.

Associate editor: Emmanuelle Lerat