Purifying Selection and Molecular Adaptation in the Genome of *Verminephrobacter*, the Heritable Symbiotic Bacteria of Earthworms

Kasper U. Kjeldsen[1,2,†], Thomas Bataillon[3,*,†], Nicolás Pinel[4], Stéphane De Mita[5], Marie B. Lund[1], Frank Panitz[6], Christian Bendixen[6], David A. Stahl[4], and Andreas Schramm[1]

[1]Department of Bioscience, Microbiology, Aarhus University, Denmark

[2]Center for Geomicrobiology, Department of Bioscience, Aarhus University, Denmark

[3]Bioinformatics Research Center, Aarhus University, Denmark

[4]Department of Civil and Environmental Engineering, University of Washington

[5]Institut de Recherche pour le Développement, Montpellier, France

[6]Department of Molecular Biology and Genetics, Aarhus University, Tjele, Denmark

†These authors contributed equally to this work.

*Corresponding author: E-mail: tbata@birc.au.dk.

Accepted: 6 February 2012

Draft (automatically annotated) genome of *Verminephrobacter aporrectodeae* subsp. *tuberculatae* strain At4^T (Vtu) genome is deposited in GenBank under accession number AFAL00000000

Abstract

While genomic erosion is common among intracellular symbionts, patterns of genome evolution in heritable extracellular endosymbionts remain elusive. We study vertically transmitted extracellular endosymbionts (*Verminephrobacter*, Betaproteobacteria) that form a beneficial, species-specific, and evolutionarily old (60–130 Myr) association with earthworms. We assembled a draft genome of *Verminephrobacter aporrectodeae* and compared it with the genomes of *Verminephrobacter eiseniae* and two nonsymbiotic close relatives (*Acidovorax*). Similar to *V. eiseniae*, the *V. aporrectodeae* genome was not markedly reduced in size and showed no A–T bias. We characterized the strength of purifying selection ($\omega = dN/dS$) and codon usage bias in 876 orthologous genes. Symbiont genomes exhibited strong purifying selection ($\omega = 0.09 \pm 0.07$), although transition to symbiosis entailed relaxation of purifying selection as evidenced by 50% higher ω values and less codon usage bias in symbiont compared with reference genomes. Relaxation was not evenly distributed among functional gene categories but was overrepresented in genes involved in signal transduction and cell envelope biogenesis. The same gene categories also harbored instances of positive selection in the *Verminephrobacter* clade. In total, positive selection was detected in 89 genes, including also genes involved in DNA metabolism, tRNA modification, and TonB-dependent iron uptake, potentially highlighting functions important in symbiosis. Our results suggest that the transition to symbiosis was accompanied by molecular adaptation, while purifying selection was only moderately relaxed, despite the evolutionary age and stability of the host association. We hypothesize that biparental transmission of symbionts and rare genetic mixing during transmission can prevent genome erosion in heritable symbionts.

Key words: symbiosis, evolution, nonsynonymous substitutions, purifying selection, positive selection, genome reduction, extracellular symbiont.

Symbiotic associations between bacteria and animal hosts shape the ecology and evolution of both partners (Dubilier et al. 2008; Moran et al. 2008; Moya et al. 2008). Transition of free-living bacteria to obligate endosymbionts profoundly affects bacterial genome evolution (Moran et al. 2008; Moya et al. 2008; Bright and Bulgheresi 2010; Toft and Andersson 2010). Periodic population bottlenecks during vertical transmission magnify the effects of genetic drift. A life cycle spent in the stable and protected host environment also reduces the scope for genetic mixing. These

Table 1
Summary of the Lifestyle and Genomic Properties of the Four Compared Species

Species	Lifestyle	Genome Size (Mbp)	Genomic G + C Content (%)	Predicted Number of Protein-Coding Genes	GenBank Accession Number
Aac	Free living	5.3	68.5	4,709	CP000512
Ajs		4.5	66.1	4,007	CP000539
Vei	Heritable symbiont	5.6	65.0	4,908	CP000542
Vtu		4.7	65.0	3,788	AFAL00000000

Data on the *Aac*, *Ajs*, and *Vei* genomes were obtained from GenBank. The *Vtu* genome size represents a minimum estimate based on the length of concatenated contigs (for details, see supplementary text, Supplementary Material online); its G + C content was calculated from a concatenation of the predicted genes.

factors lessen the efficacy of natural selection and explain genomic erosion, that is, accelerated substitution rates, biased nucleotide base composition, pseudogenization, and gene loss in symbiont genomes. Genome erosion is best documented for obligate intracellular endosymbionts of insects (Moran et al. 2008, 2009; Toft and Andersson 2010) and has become the paradigm for genome evolution of such heritable symbionts (Andersson and Kurland 1998). In contrast, little (and conflicting) information is available for extracellular bacterial endosymbionts (Dubilier et al. 2008); while those of stinkbugs show vertical transmission, cospeciation, and reductive genome evolution in line with intracellular insect endosymbionts (Hosokawa et al. 2006; Kikuchi et al. 2009), the genomes of gutless oligochaete symbionts are similar in size and G + C content to their free-living relatives; however, their mode of transmission has not been fully resolved (Woyke et al. 2006).

The genome of the extracellular earthworm endosymbiont *Verminephrobacter eiseniae* (Pinel et al. 2008) shows no signs of genome reduction either (Pinel 2009). *Verminephrobacter* (Betaproteobacteria) are species specific, extracellular endosymbionts of most lumbricid earthworm species (Lund et al. 2010). They colonize the earthworm's excretion organs, nephridia (Schramm et al. 2003), and are transmitted vertically: symbionts are deposited along with eggs and sperm in the egg capsules (Davidson and Stahl 2006) and during embryogenesis, selectively recruited by the earthworm (Davidson and Stahl 2008). Fitness experiments have shown that the symbionts increase the reproductive success of their host (Lund, Holmstrup, et al. 2010); the earthworm–*Verminephrobacter* symbiosis is therefore considered mutually beneficial. Although host and symbiont can be separated under laboratory conditions (Lund, Holmstrup, et al. 2010), the host-associated lifestyle of *Verminephrobacter* is evolutionary stable and as old as that of many heritable intracellular endosymbionts of insects (Moran et al. 2008). Strict host fidelity and a pattern of cospeciation indicate an origin of the *Verminephrobacter*–earthworm symbiosis in the most recent common ancestor of lumbricid earthworms, 60–130 Myr (Lund et al. 2010).

Here, we obtain a draft genome from a second *Verminephrobacter* species (Lund et al. 2012). We contrast the strength of purifying selection and the level of positive selection in the two *Verminephrobacter* genomes with those found in two free-living bacteria from the sister genus *Acidovorax* to study how the ancient transition to a symbiotic lifestyle affected genome evolution in these extracellular heritable endosymbionts.

Lack of Genome Erosion in *Verminephrobacter* Symbionts

The draft genome assembly of *Verminephrobacter aporrectodeae* subsp. *tuberculatae* strain At4^T (*Vtu*) consisted of 1,082 unique contigs with an average read depth of 45× (see Materials and Methods and supplementary text, Supplementary Material online). The *Vtu* and the *V. eiseniae* strain EF01-2^T (*Vei*) genomes are similar in size, gene, and G + C content to those of *Acidovorax avena* subsp. *citrulli* strain AAC00-1 (*Aac*) and *Acidovorax* sp. JS42 (*Ajs*) (table 1). By comparing the gene complement encoded by the *Vtu* genome against those of *Vei*, *Aac*, and *Ajs*, it is obvious that the size of the *Vtu* draft genome assembly (4.6 Mbp) underestimates the actual genome size, which is likely close to the size of the *Vei* genome (supplementary text, Supplementary Material online). Thus the symbiont genomes have apparently escaped the size reduction and strong A–T bias typical of heritable symbiotic bacteria. Using reciprocal basic local alignment search tool (Blast), we identified a set of 1,038 orthologous genes shared among the four genomes. Phylogenetic analyses of various taxonomic marker genes consistently group all known *Verminephrobacter* species in a monophyletic clade with *Acidovorax* as a sister lineage (Pinel et al. 2008; Lund et al. 2010). We inferred by maximum likelihood, a four-species unrooted phylogeny for each ortholog alignment. In 961 cases out of the 1,038, the species tree topology described above was strongly supported by an Akaike's Information Criterion (AIC) at least three units lower (better) than alternative topologies. The remaining 77 orthologs were discarded from further analyses because they did not discriminate reliably between topologies (n = 65) or supported alternative topologies (n = 12). To further guard against spurious results due to saturation of nucleotide substitutions, we examined the total number of synonymous substitutions estimated on each ortholog

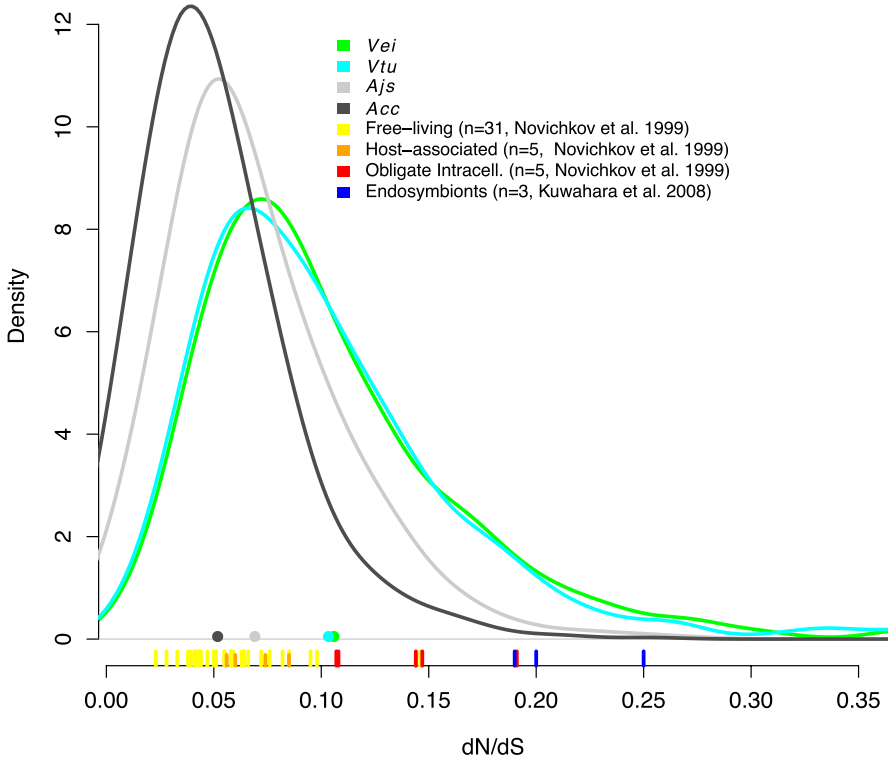

Fig. 1.—Intensity of purifying selection (d*N*/d*S*) in symbiont (*Vei*, *Vtu*) and in nonsymbiont (*Aac*, *Ajs*) genomes. Distribution of d*N*/d*S* in each species was inferred from the least diverged data set ($n = 876$ orthologs) and assuming branch model H1 (see supplementary fig. S1, Supplementary Material online). The empirical frequency distributions obtained for each species were smoothed using a Gaussian kernel method. Note that a few d*N*/d*S* values larger than 1 were discarded before smoothing (3–7 data points excluded per genome). Genome average d*N*/d*S* estimates for each species are also indicated (colored bullet). Data from the literature are depicted in the lower part of the plot (vertical ticks) for comparison. We used reports of 44 average d*N*/d*S* values based on at least 300 orthologs per set of genomes and classified genomes according to their lifestyle (see color legend).

phylogeny (supplementary fig. S2, Supplementary Material online) and discarded every ortholog exhibiting >6 synonymous substitutions expected per site (summed over the entire tree). This left a set of $n = 876$ orthologs representing "least diverged" genes conforming to the species tree.

With this set, we quantified divergence between the two symbiont genomes. Pairwise divergence between the *Vei* and *Vtu* genomes accounts for about one-third of the total divergence in each of the unrooted ortholog phylogenies with 1 and 0.1 expected synonymous and nonsynonymous substitution per site, respectively (mean d*S* ± standard deviation [SD] = 1.05 ± 0.33; mean d*N* ± SD = 0.1 ± 0.05; supplementary fig. S2 and table S6, Supplementary Material online). We then quantified selective constraints in the symbiont genomes by estimating the ratio of nonsynonymous to synonymous substitution rates $\omega = dN/dS$. Values of $\omega < 1$ indicate purifying selection, that is, most amino acid substitutions are deleterious and thus removed, $\omega = 1$ indicates the absence of constraints, whereas $\omega > 1$ indicates that fixation of amino acid–replacing mutations is driven by positive selection. We estimated ω on each terminal branch of the unrooted ortholog phylogenies (fig.1, branch model H1, supplementary fig. S1, Supplementary Material online). Median ω values were 0.04 and 0.06 in the *Aac* and *Ajs* genomes, respectively, and 0.09 in both symbiont genomes. These values are within the range reported for free-living bacteria (0.02–0.09; fig. 1; Novichkov et al. 2009) and lower than in heritable endosymbionts and host-associated pathogens (0.11–0.25; fig. 1; Kuwahara et al. 2008; Kuo et al. 2009; Novichkov et al. 2009), indicating that *Verminephrobacter* genomes still undergo fairly strong purifying selection.

Relaxed Purifying Selection in *Verminephrobacter* Genomes Compared with Their Free-Living Relatives

Despite substantial purifying selection in *Verminephrobacter*, we found clear evidence for relaxed purifying selection in the *Vei* and *Vtu* genomes relative to *Acidovorax*. The genome-wide distribution of ω was shifted toward higher

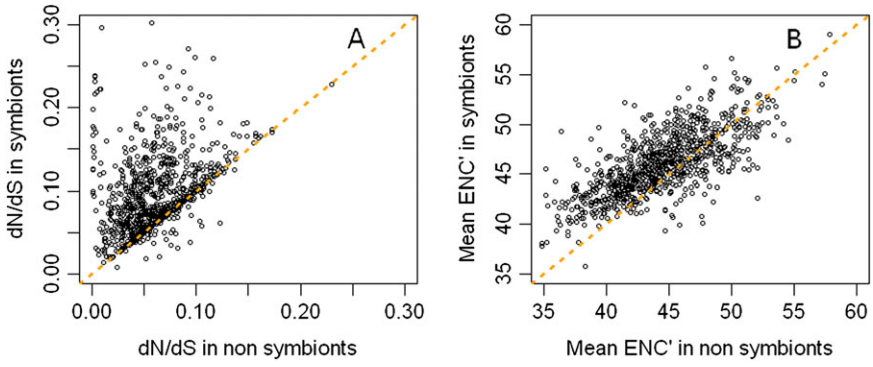

Fig. 2.—Genome-wide relaxation of purifying selection in symbiont genomes. (*A*) Comparison of the selective constraint (d*N*/d*S*) in symbiont (*Vei*, *Vtu*) versus nonsymbiont (*Aac*, *Ajs*) branches of unrooted ortholog phylogenies. Each point represents the d*N*/d*S* value for each ortholog from the least diverged data set ($n = 876$). Values of d*N*/d*S* reported here are obtained using a model averaging procedure over the set of branch models (H0–H4; for details, see supplementary fig. S1 and text, Supplementary Material online). (*B*) Comparison of the intensity of codon usage bias (ENCp) in symbiont versus nonsymbiont genomes. The intensity of codon usage for each ortholog was computed as the average of the ENCp estimate for symbionts and nonsymbionts.

values in the symbiont genomes (fig. 1), and 755 of the 876 orthologs had a higher ω (binomial test, $P < 10^{-10}$) in the symbiont versus nonsymbiont terminal branches (fig. 2*A*). This finding is robust on three accounts. First, our analysis is based on a model-averaging procedure to ensure reliable estimation of ω (see Materials and Methods). Second, the finding is not an artifact of a single symbiont genome (*Vtu*) being poorly sequenced and yielding apparent elevated nonsynonymous substitution rates (supplementary fig. S3, Supplementary Material online). Last, even when restricting our analysis to subsets of orthologs with $dS < 2.5$ (corresponding to ~0.5 expected substitution per branch and site), the vast majority of orthologs exhibited inflated ω values on the symbiont branches (supplementary fig. S4, Supplementary Material online).

Codon usage bias analysis represents an independent way to document relaxed purifying selection (Hershberg and Petrov 2008). Less bias is expected when purifying selection fails to remove slightly deleterious mutations with suboptimal codon usage. We quantified codon usage bias in the four species through the effective number of codons (ENCp) used (Novembre 2002). ENCp is inversely proportional to the intensity of codon usage bias; high values therefore indicate relaxed purifying selection. In *Vei* and *Vtu*, 90% of the 876 orthologs had higher ENCp values (binomial test, $P < 10^{-10}$) (fig. 2*B*), providing additional evidence for relaxed purifying selection in the *Verminephrobacter* genomes.

The degree of relaxation in purifying selection varies across different functional gene categories; this has been used to infer the role of genes in insect endosymbionts (Toft et al. 2009; Pérez-Brocal et al. 2011): greatest relaxation is expected for genes rendered obsolete upon transition to symbiosis, while strong purifying selection indicates essential gene functions. Querying for orthologs with the highest absolute increase in ω ($\Delta\omega > 0.09$) between symbiont and nonsymbiont lineages, we identified 102 orthologs that experienced the greatest relaxation of purifying selection in the two symbiont genomes (supplementary table S1, Supplementary Material online). To ensure that large ω values are not caused by erroneous recruitment of paralogs, we used BlastP searches to identify the number of per-genome paralogs for each of the 102 orthologs. More than 90 orthologs had no identifiable paralogs in their parent genomes (supplementary table S1, Supplementary Material online). In the case of paralogs present, we validated inferred orthologies by phylogenetic reconstruction (supplementary fig. S6, Supplementary Material online). This analysis left 100 orthologs that experienced the strongest relaxation of purifying selection in the symbiont lineages. We examined the predicted functional categories of these orthologs using Clusters of Orthologous Groups (COG)-based analysis and annotation (Tatusov et al. 2001) and found significant overrepresentation of COG categories T (signal transduction; binomial test, $P < 0.002$) and M (cell envelope biogenesis/outer membrane functions, $P < 0.00005$) compared with the set of 876 orthologs (fig. 3*A*). This suggests that a sizeable proportion of the genes involved in signal transduction and surface structures became dispensable when the *Verminephrobacter* lineage became symbiotic. Such genes govern interactions with the environment and likely experience dramatic changes in selective pressure upon transition from a free-living to a symbiotic lifestyle; genes in the latter category are often disproportionately lost in heritable intracellular symbionts (Toft and Andersson 2010). In contrast, orthologs involved in translation (COG category J) and most notably in membrane transport (COG categories E, F, G, and P) were about 3-fold underrepresented ($P < 0.0005$), suggesting a key importance of uptake systems for the *Verminephrobacter*–earthworm symbiosis. Few instances of strongly relaxed purifying selection occur in the otherwise abundant COG category E (amino acid transport and metabolism, supplementary fig. S5, Supplementary Material

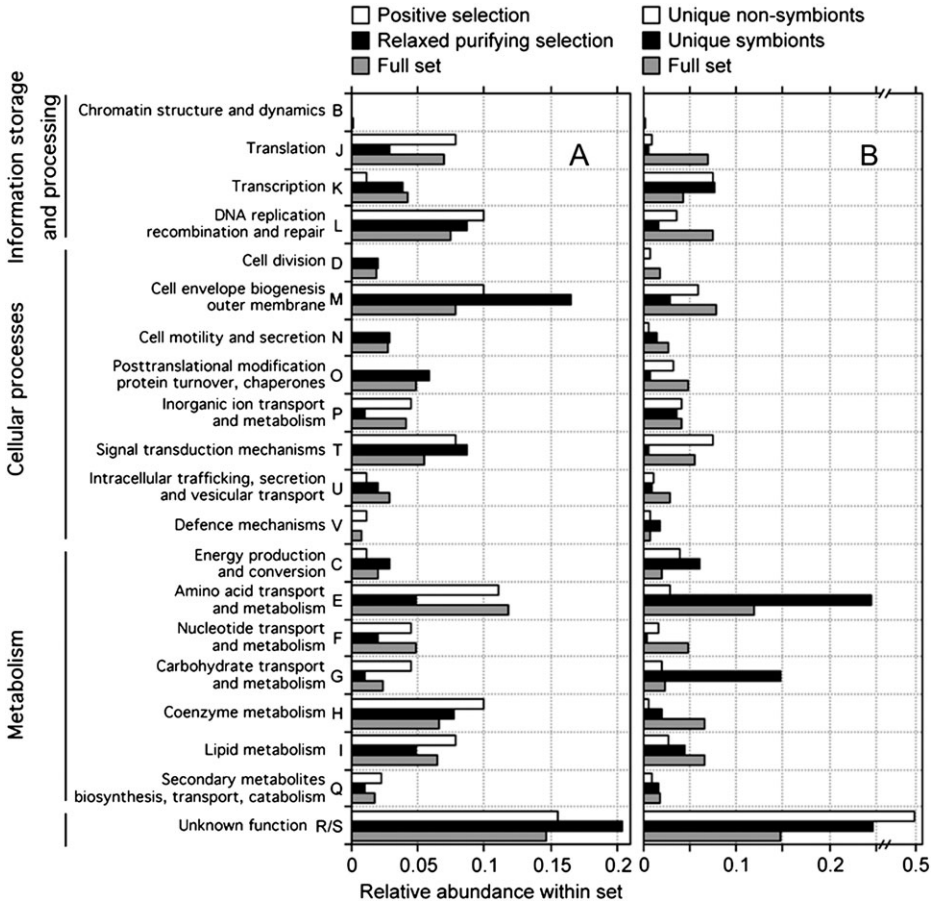

Fig. 3.—COG-based profiling of gene function. (*A*) Profiles of the least diverged set of orthologs shared by all four genomes ($n = 876$) and profiles of the subsets containing either instances of greatest relaxed purifying selection ($n = 100$) or most significant instances of positive selection in the symbiotic branches (*Vei*, *Vtu*; $n = 89$). (*B*) Profiles of genes unique to the nonsymbiont (*Aac*, *Ajs*; $n = 441$) or the symbiont genomes ($n = 975$) compared with the least diverged set of orthologs.

online). This pattern is consistent with the hypothesized symbiont function in nitrogen recycling for their host (Pandazis 1931; Schramm et al. 2003). Manual annotation of the 100 orthologs under most relaxed purifying selection in *Vei* and *Vtu* confirmed the COG-based analysis (supplementary table S1, Supplementary Material online). Orthologs involved in biosynthesis of membrane lipoproteins and glycolipids, and in biosynthesis and degradation of peptidoglycan represented COG category M, while two-component systems and transcriptional regulators represented COG category T (supplementary table S1, Supplementary Material online).

Footprints of Molecular Adaptation in the Symbiont Genomes

Transition to symbiosis can be seen as colonization of a new ecological niche (the host) triggering episodes of adaptation detectable at the molecular level (Toft and Andersson 2010). We used a conservative branch-site likelihood ratio test (Zhang et al. 2005) to identify orthologs with sites evolving under positive selection ($\omega > 1$) in the symbiont terminal branches of the four-species phylogeny. To account for multiple testing ($n = 876$ tests), we used a false discovery rate (FDR) approach and detected 90 orthologs evolving under positive selection in the symbiont genomes ($P < 0.005$, estimated FDR < 0.03 among significant tests). Using the same threshold, we only detected 10 orthologs evolving under positive selection in the nonsymbiont genomes. Scrutinizing the 90 ortholog sets for erroneous recruitment of paralogs (see above), we identified one set where paralogy may bias our inference (supplementary fig. S7, Supplementary Material online), leaving a total of 89 orthologs of which 68 have no paralogs in any of the four genomes (supplementary table S2, Supplementary Material online). Another concern is the quality of the draft genome (454 sequencing) data. However, for the 89 positively selected genes 1) the overall d*N*/d*S* values for *Vtu* and *Vei* are similar (supplementary fig. S3, Supplementary Material online), 2) rates of homopolymeric runs are not systematically higher in *Vtu*

relative to *Vei* or *Acidovorax* (supplementary table S8, Supplementary Material online), 3) gene length is virtually identical in *Vtu* and *Vei* (supplementary table S8, Supplementary Material online), and 4) frame shifts are absent; we are thus confident that detection of positive selection is not biased by 454 sequencing errors.

The 89 orthologs were functionally broadly distributed, with COG categories T, M, L and H marginally overrepresented relative to the initial set of orthologs (fig. 3*A*). As discussed above, genes involved in signal transduction (T) and outer membrane functions (M) govern interactions with the environment. While some of these genes may become obsolete upon transition to a symbiotic lifestyle, others may become essential for symbiont-host interactions and undergo adaptive evolution. Fifteen of the 89 orthologs were also in the subset of 100 orthologs showing greatest relaxation of purifying selection (supplementary tables S1 and S2, Supplementary Material online). Two nonexclusive processes can explain the overlap. First, strong positive selection on a gene will increase its average d*N*/d*S* in the symbionts clade possibly lumping it in the "most relaxed selection" category. Second, a gene could experience relaxed purifying selection after transition to symbiosis, which, in turn, may open new avenues for adaptation.

The 89 orthologs were manually annotated (supplementary table S2, Supplementary Material online). Orthologs with a predicted function in signal transduction (T) primarily comprised two-component systems, which have a key role in niche adaptation of bacteria (Alm et al. 2006). Orthologs with a predicted outer membrane function represented lipid A-synthase (set 923), lipid A-sugar transferases (sets 333, 339), and lipoprotein acyl-transferases (sets 93, 95). Such genes often undergo positive selection in pathogenic bacteria, where they are important, for example, for attachment and evasion of host defenses (Fitzpatrick et al. 2005; Chen et al. 2006; Soyer et al. 2009). Apolipoprotein acyl-transferase Lnt (set 93) evolves under positive selection in uropathogenic *Escherichia coli* that, like *Verminephrobacter*, colonize a urinary tract environment (Chen et al. 2006). In *Vibrio fischeri*, the extracellular symbiont of the *Euprymna* squid light organ, lipid A-acylating enzymes are key players in colonization (Adin et al. 2008). One of the lipid A-acylating proteins of *V. fischeri* (HtrB1, locus tag VF_A0687) is homologous (30% pairwise amino acid identity) to ortholog set 339. Like *V. fischeri*, *Verminephrobacter* colonize their hosts by recruitment from a mixed microbial community, migration to a specific host tissue, and attachment (Davidson and Stahl 2008). Surface lipoproteins and glycolipids may thus be important for both symbiont–host recognition and for attachment.

Several of the 89 orthologs inferred to undergo positive selection have a predicted function in peptidoglycan biosynthesis (sets 297, 304) and degradation (sets 155, 306, 326, 688, and 1477) (supplementary table S2, Supplementary Material online). Peptidoglycan monomers mediate host interactions of pathogenic and symbiotic bacteria and are recognized by the innate immune system of animals (Cloud-Hansen et al. 2006, Moya et al. 2008). Accordingly, N-acetylmuramoyl-L-alanine amidase, a putative homolog of ortholog set 326, also undergoes positive selection in uropathogenic *E. coli* (Chen et al. 2006). Peptidoglycan constituents are also important for the *V. fischeri–Euprymna* symbiosis (Adin et al. 2009, Altura et al. 2011). The *V. fischeri* membrane transport protein AmpG (locus tag VF_0720) controls the extracellular concentration of peptidoglycan monomers released by hydrolytic enzymes (like orthologous set 306 and 326), which in turn affects the morphogenesis of the *Euprymna* light organ. The *Vei* and *Vtu* orthologs of set 1386 (supplementary table S1, Supplementary Material online) are AmpG homologs sharing 32% full-length amino acid identity with the *V. fischeri* AmpG and may serve a similar function as AmpG in *V. fischeri*. Further discussion of the functional significance of the orthologs evolving under positive selection in *Vei* and *Vtu* is included in the footnote of supplementary table S2, Supplementary Material online.

Drivers for Genome Evolution in Extracellular Heritable Symbionts

Unlike intracellular insect endosymbionts (Moran et al. 2008) and stinkbug extracellular endosymbionts (Hosokawa et al. 2006; Kikuchi et al. 2009), the earthworm symbiont genomes are not eroded, despite their ancient association, host fidelity, and vertical transmission (Davidson and Stahl 2006, 2008; Lund et al. 2010). Lack of genomic synteny between *Vei* and *Aac/Ajs* and a high number of mobile elements in the *Vei* genome (Pinel 2009) may indicate that *Verminephrobacter* moves toward genome reduction according to the symbiont stage theory (Moran et al. 2009; Toft and Andersson 2010). However, age of the symbiosis (60–130 Myr), strong purifying selection, and low levels of pseudogenization (Pinel 2009) do not support this scenario. Furthermore, analysis of genes unique for the symbiont lineage (fig. 3*B*) suggests that *Verminephrobacter* have acquired novel genes after transition into symbiosis, most prominently coding for amino acid and carbohydrate transport, while genes coding for signal transduction are underrepresented compared with the reference genomes.

Our results show a more nuanced pattern of genome evolution in *Verminephrobacter* compared with heritable intracellular symbionts of insects, that is, the absence of genome reduction and A–T bias, fairly strong purifying selection, pervasive molecular adaptation, and signs of gene acquisition after transition to symbiosis. We propose that this pattern of genome evolution is best explained by the life cycle of the symbionts (Bright and Bulgheresi 2010). *Verminephrobacter* experience a short free-living stage during vertical transmission: the symbionts move to the worm

surface and are, together with other bacteria, enclosed in egg capsules, where they colonize the developing embryos (Davidson and Stahl 2006, 2008). It is possible that earthworms exchange symbionts residing in their mucus during mating, leading to biparental inheritance. This mode of transmission may select against the loss of genes essential for surviving outside the host and may offer a window of opportunity for recombination and lateral gene transfer, which will slow down or prevent Muller's ratchet (Felsenstein 1974; Moran 1996). In contrast, vertical transmission of insect intracellular and stinkbug extracellular endosymbionts is uniparental (Hosokawa et al. 2006; Moran et al. 2008) and stringent, effectively preventing genetic mixing.

Material and Methods

Organisms and Genome Sequencing

The annotated genome sequences of *Aac*, *Ajs*, and *Vei* were downloaded from the NCBI website (table 1). A draft genome sequence of *Vtu* was obtained by pyrosequencing using the GS20 and GS20 FLX platforms (Roche 454 Life Sciences) as per the manufacturers recommendations. Genomic DNA was extracted from a *Vtu* culture grown in liquid R2A medium using the "DNeasy Blood and Tissue Kit" (Qiagen). A shotgun library was prepared from 50 μg DNA using the "Standard DNA Library Preparation Kit" (Roche). Nebulized, purified, and A and B adapter-linked single strand DNA fragments (average size 200–700 bp) were clonally amplified using the "Emulsion PCR Kit I" (Roche). Four independent GS20 runs and one GS20 FLX sequencing run were performed to produce 1,625,984 reads with an average length of 96 bp (GS20) and 149,562 reads with an average length of 242 bp (GS20 FLX), together totaling 192,595,373 sequenced bases. De novo sequence assembly was performed using CLC Genomics Workbench (version 4.0.2) generating 1,082 unique contigs (size range 200–81,277 bp, average size 4,326 bp) comprising 4,681,801 bp. The *Vtu* draft genome sequence was assigned accession number AFAL00000000 in the GenBank WGS sequence repository.

Putative protein-coding genes were identified with GLIMMER v3.02 (Delcher et al. 2007) using default settings and a lower size cutoff of 150 bp. Predicted genes were validated by homology searches using inferred amino acid sequences as Blast search queries against release 45 of the NCBI refseq protein database (Pruitt et al. 2007). Hits with an e-value cutoff $<1 \times 10^{-20}$ were considered significant, leaving 3,788 predicted protein-coding genes (table 1).

Identification of Orthologous Gene Sets

Sets of orthologous genes shared by the four species were identified via amino acid sequence-based reciprocal Blast searches using an e-value cutoff of 1×10^{-20}. Inferred amino acid sequences of each identified gene set were aligned using ClustalW v1.81 (Thompson et al. 1994) and those sets comprised of members with less than 20% difference in sequence length and that shared >35% sequence identity across their entire length were considered orthologous. DNA sequence-based codon alignments of the orthologous gene sets were constructed from the corresponding aligned amino acid sequences using PAL2NAL v12.2 (Suyama et al. 2006). For functional assignment, orthologs were manually compared against InterPro (Zdobnov and Apweiler 2001) and COG (Tatusov et al. 2001) databases. Automated functional assignment was performed by rps-Blast (Marchler-Bauer et al. 2002) (standalone Blast release 2.2.24) against the preformatted COG v.1.0 database. The top e-value COG match was recorded for each query.

Phylogenetic and Evolutionary Analyses

A phylogenetic tree was inferred for each set of orthologs. As all sets comprise four species, three unrooted tree topologies are possible: T1: (*Aac*, *Ajs*, [*Vei*, *Vtu*]), T2: (*Aac*, *Vei*, [*Ajs*, *Vtu*]), and T3: (*Aac*, *Vtu*, [*Ajs*, *Vei*]). T1 represents the canonical species trees. The likelihood of each topology was obtained under the general time reversible DNA substitution model using PhyML (Guindon and Gascuel 2003). AIC was used to compare the fit of alternative topologies. For each of the least diverged orthologs ($n = 876$), a set of alternative "branch models" (supplementary fig. S1, Supplementary Material online) was used to estimate ω and model changes in ω along the underlying topology (here fixed to T1). The *codeml* program in the PAML package (wrapped using python scripts) was used to estimate jointly ω and the transition to transversion ratio using the "F3×4" model (Yang 2007). A model-averaging procedure was used for robust estimation of ω values in the symbiont ("V") versus nonsymbiont ("A") terminal branches of the T1 phylogeny. Briefly, for each ortholog, we obtained the AIC of 5 branch models (supplementary fig. S1, Supplementary Material online) as $\mathrm{AIC} = -2 \ln L + 2p$, where $\ln L$ denotes the log-likelihood of the data under the model and p the number of parameters fitted. We used differences in AIC ($\Delta\mathrm{AIC}_i$) among the set of alternative branch models {H0, H1, H2, H3, H4} to calculate a set of weights $\{z_0, z_1, z_2, z_3, z_4\}$ as $z_i = \mathrm{Exp}(\Delta\mathrm{AIC}i)/[\Sigma i\ \mathrm{Exp}(\Delta\mathrm{AIC}_i)]$, where summation is over the set of five models. We then estimated ω_A (respectively ω_V) as the weighted average of ω values under each model (using z's as weights). This procedure guards against sensitivity of estimation of ω values to a single branch model. We used the effective number of codons (ENC) and its extension to account for heterogeneity in nucleotide composition, ENC prime (ENCp), to quantify bias in codon usage (Novembre 2002). ENCp values were estimated for each gene and species separately using the software ENCprime (version February 2006; Novembre 2002). Whenever needed, we

controlled for the FDR (Storey and Tibshirani 2003). All statistical analyses and model-averaging calculations were carried out in R (http://www.R-project.org).

Supplementary Material

Supplementary text, figures S1–S5, and tables S1–S8 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

Funding for genome sequencing was provided by the Carlsberg Foundation, Denmark. The Danish Research Council supported T.B. and K.U.K. We thank Sten Andersen for continuous and invaluable help with maintaining our computer server.

Literature Cited

Adin DM, Engle JT, Goldman WE, McFall-Ngai MJ, Stabb EV. 2009. Mutations in *ampG* and lytic transglycosylase genes affect the net release of peptidoglycan monomers from *Vibrio fischeri*. J Bacteriol. 191:2012–2022.

Adin DM, et al. 2008. Characterization of *htrB* and *msbB* mutants of the light organ symbiont *Vibrio fischeri*. Appl Environ Microbiol. 74:633–644.

Alm E, Huang K, Arkin A. 2006. The evolution of two-component systems in bacteria reveals different strategies for niche adaptation. PLoS Comput Biol. 2:e143.

Altura MA, Stabb E, Goldman W, Apicella M, McFall-Ngai MJ. 2011. Attenuation of host NO production by MAMPs potentiates development of the host in the squid-vibrio symbiosis. Cell Microbiol. 13:527–537.

Andersson SG, Kurland CG. 1998. Reductive evolution of resident genomes. Trends Microbiol. 6:263–268.

Bright M, Bulgheresi S. 2010. A complex journey: transmission of microbial symbionts. Nat Rev Microbiol. 8:218–230.

Chen SL, et al. 2006. Identification of genes subject to positive selection in uropathogenic strains of *Escherichia coli*: a comparative genomics approach. Proc Natl Acad Sci U S A. 103:5977–5982.

Cloud-Hansen KA, et al. 2006. Breaching the great wall: peptidoglycan and microbial interactions. Nat Rev Microbiol. 4:710–716.

Davidson SK, Stahl DA. 2006. Transmission of nephridial bacteria of the earthworm *Eisenia fetida*. Appl Environ Microbiol. 72:769–775.

Davidson SK, Stahl DA. 2008. Selective recruitment of bacteria during embryogenesis of an earthworm. ISME J. 2:510–518.

Delcher AL, Bratke KA, Powers EC, Salzberg SL. 2007. Identifying bacterial genes and endosymbiont DNA with Glimmer. Bioinformatics 23:673–679.

Dubilier N, Bergin C, Lott C. 2008. Symbiotic diversity in marine animals: the art of harnessing chemosynthesis. Nat Rev Microbiol. 6:725–740.

Felsenstein J. 1974. The evolutionary advantage of recombination. Genetics 78:737–756.

Fitzpatrick DA, Creevey CJ, McInerney JO. 2005. Evidence of positive Darwinian selection in putative meningococcal vaccine antigens. J Mol Evol. 61:90–98.

Guindon S, Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol. 52:696–704.

Hershberg R, Petrov DA. 2008. Selection on codon bias. Annu Rev Genet. 42:287–299.

Hosokawa T, Kikuchi Y, Nikoh N, Shimada M, Fukatsu T. 2006. Strict host-symbiont cospeciation and reductive genome evolution in insect gut bacteria. PLoS Biol. 4:e337.

Kikuchi Y, et al. 2009. Host-symbiont co-speciation and reductive genome evolution in gut symbiotic bacteria of acanthosomatid stinkbugs. BMC Biol. 7:2.

Kuo CH, Moran NA, Ochman H. 2009. The consequences of genetic drift for bacterial genome complexity. Genome Res. 19:1450–1454.

Kuwahara H, et al. 2008. Reductive genome evolution in chemoautotrophic intracellular symbionts of deep-sea *Calyptogena* clams. Extremophiles 12:365–374.

Lund MB, et al. 2010. Diversity and host specificity of the *Verminephrobacter*-earthworm symbiosis. Environ Microbiol. 12:2142–2151.

Lund MB, Holmstrup M, Lomstein BA, Damgaard C, Schramm A. 2010. Beneficial effect of *Verminephrobacter* nephridial symbionts on the fitness of the earthworm *Aporrectodea tuberculata*. Appl Environ Microbiol. 76:4738–4743.

Lund MB, Schätzle S, Schramm A, Kjeldsen KU. 2012. *Verminephrobacter aporrectodeae* sp. nov. subsp. *tuberculatae* and subsp. *caliginosae*; the specific nephridial symbionts of the earthworms *Aporrectodea tuberculata* and *A. caliginosa*. Antonie Van Leeuwenhoek. 101:507–514.

Marchler-Bauer A, et al. 2002. CDD: a database of conserved domain alignments with links to domain three-dimensional structure. Nucleic Acids Res. 30:281–283.

Moran NA. 1996. Accelerated evolution and Muller's rachet in endosymbiotic bacteria. Proc Natl Acad Sci U S A. 93:2873–2878.

Moran NA, McCutcheon JP, Nakabachi A. 2008. Genomics and evolution of heritable bacterial symbionts. Annu Rev Genet. 42:165–190.

Moran NA, McLaughlin HJ, Sorek R. 2009. The dynamics and time scale of ongoing genomic erosion in symbiotic bacteria. Science 323:379–382.

Moya A, Pereto J, Gil R, Latorre A. 2008. Learning how to live together: genomic insights into prokaryote-animal symbioses. Nat Rev Genet. 9:218–229.

Novembre JA. 2002. Accounting for background nucleotide composition when measuring codon usage bias. Mol Biol Evol. 19:1390–1394.

Novichkov PS, Wolf YI, Dubchak I, Koonin EV. 2009. Trends in prokaryotic evolution revealed by comparison of closely related bacterial and archaeal genomes. J Bacteriol. 191:65–73.

Pandazis G. 1931. Zur Frage der Bakteriensymbiose bei Oligochaeten. Zentralbl Bakteriol. 120:440–453.

Pérez-Brocal V, Gil R, Moya A, Latorre A. 2011. New insights on the evolutionary history of aphids and their primary endosymbiont *Buchnera aphidicola*. Int J Evol Biol. 2011:250154.

Pinel N. 2009. Physiological and genomic insights into the biology of *Verminephrobacter eiseniae*, a bacterial symbiont of the earthworm *Eisenia fetida* [PhD dissertation]. [Seattle (WA)]: University of Washington.

Pinel N, Davidson SK, Stahl DA. 2008. *Verminephrobacter eiseniae* gen. nov., sp. nov., a nephridial symbiont of the earthworm *Eisenia foetida* (Savigny). Int J Syst Evol Microbiol. 58:2147–2157.

Pruitt KD, Tatusova T, Maglott DR. 2007. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res. 35:D61–D65.

Schramm A, et al. 2003. *Acidovorax*-like symbionts in the nephridia of earthworms. Environ Microbiol. 5:804–809.

Soyer Y, Orsi RH, Rodriguez-Rivera LD, Sun Q, Wiedmann M. 2009. Genome wide evolutionary analyses reveal serotype specific patterns of positive selection in selected *Salmonella* serotypes. BMC Evol Biol. 9:264.

Storey JD, Tibshirani R. 2003. Statistical significance for genomewide studies. Proc Natl Acad Sci U S A. 100:9440–9445.

Suyama M, Torrents D, Bork P. 2006. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Res. 34:W609–W612.

Tatusov RL, et al. 2001. The COG database: new developments in phylogenetic classification of proteins from complete genomes. Nucleic Acids Res. 29:22–28.

Thompson JD, Higgins DG, Gibson TJ. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22:4673–4680.

Toft C, Andersson SG. 2010. Evolutionary microbial genomics: insights into bacterial host adaptation. Nat Rev Genet. 11:465–475.

Toft C, Williams TA, Fares MA. 2009. Genome-wide functional divergence after the symbiosis of proteobacteria with insects unraveled through a novel computational approach. PLoS Comput Biol. 5:e1000344.

Woyke T, et al. 2006. Symbiosis insights through metagenomic analysis of a microbial consortium. Nature 443:950–955.

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

Zdobnov EM, Apweiler R. 2001. InterProScan—an integration platform for the signature-recognition methods in InterPro. Bioinformatics 17:847–848.

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.

Associate editor: Brandon Gaut