Hamiltonella defensa, genome evolution of protective bacterial endosymbiont from pathogenic ancestors

Patrick H. Degnana,b,1, Yeisoo Yub, Nicholas Sisnerosb, Rod A. Wingb, and Nancy A. Morana

aDepartment of Ecology and Evolutionary Biology, bArizona Genomics Institute, University of Arizona, Tucson, AZ 85721

Edited by Edward F. DeLong, Massachusetts Institute of Technology, Cambridge, MA, and approved April 14, 2009 (received for review January 7, 2009)

Eukaryotes engage in a multitude of beneficial and deleterious interactions with bacteria. Hamiltonella defensa, an endosymbiont of aphids and other sap-feeding insects, protects its aphid host from attack by parasitoid wasps. Thus H. defensa is only conditionally beneficial to hosts, unlike ancient nutritional symbionts, such as Buchnera, that are obligate. Similar to pathogenic bacteria, H. defensa is able to invade naive hosts and circumvent host immune responses. We have sequenced the genome of H. defensa to identify possible mechanisms that underlie its persistence in healthy aphids and protection from parasitoids. The 2.1-Mb genome has undergone significant reduction in size relative to its closest free-living relatives, which include Yersinia and Serratia species (4.6–5.4 Mb). Auxotrophic for 8 of the 10 essential amino acids, H. defensa is reliant upon the essential amino acids produced by Buchnera. Despite these losses, the H. defensa genome retains more genes and pathways for a variety of cell structures and processes than do obligate symbionts, such as Buchnera. Furthermore, putative pathogenicity loci, encoding type-3 secretion systems, and toxin homologs, which are absent in obligate symbionts, are abundant in the H. defensa genome, as are regulatory genes that likely control the timing of their expression. The genome is also littered with mobile DNA, including phage-derived genes, plasmids, and insertion-sequence elements, highlighting its dynamic nature and the continued role horizontal gene transfer plays in shaping it.

Insects host a wide diversity of noncultivable bacteria, which have important ecological phenotypes ranging from parasitism to mutualism (1, 2). Genome sequencing of noncultivable parasitic bacteria has revealed possible mechanisms responsible for reproductive manipulations (3–5), whereas genomes of obligate mutualists of ants, aphids, psyllids, and tsetse flies, and sharpshooters have documented biosynthetic abilities important to host nutrition (6–10). Heritable endosymbionts that protect their hosts from parasites and pathogens are increasingly being recognized as common. Because they are occasionally transferred horizontally, sometimes between distantly related species, these symbionts provide a conduit for the transfer of highly adaptive and stably inherited traits (resistance and defense) between host species. So far, no such defensive symbiont has been studied using genome sequencing.

Hamiltonella defensa, a gamma-proteobacterium, is a maternally transmitted defensive endosymbiont found sporadically in sap-feeding insects, including aphids, psyllids, and whiteflies (11–13). In pea aphids (Acyrthosiphon pisum), H. defensa can block larval development of the solitary endoparasitoid wasps Aphidius ervi and Aphidius cadyi, rescuing the aphid host (14–16). The reduction in aphid mortality is variable among H. defensa strains and is correlated to the presence of a temperate, lambda-like bacteriophage APSE, which infects H. defensa (17–20). H. defensa occurs sporadically in A. pisum and is beneficial only when parasitoids are present (21). Consequently, infection frequencies increase under strong parasitoid pressure but decrease when parasitoids are absent. H. defensa and APSE can also be transmitted horizontally either intraspecifically [e.g., sexually (22)] or interspecifically (12, 17). Moreover, protection by H. defensa has been shown to be transferable between distantly related aphid species (19).

Although H. defensa confers protection, it also exhibits many attributes of enteric pathogens. Its lifestyle requires that it invade novel hosts, and a preliminary survey of its genome content showed that it contains many pathogenicity factors related to host invasion (18). APSE strains encode toxins, including cytolysins, distending toxin and Shiga-like toxin, intimating a role of horizontal gene transfer (HGT) in modulating the protection conferred by H. defensa (18, 23).

To shed light on the interactions of H. defensa, its insect hosts, bacteriophage, and invading parasitoids, we have sequenced the H. defensa genome from a strain previously shown to confer protection to A. pisum (16). The H. defensa genome combines mechanisms known from both symbiotic and pathogenic bacterial species.

Results and Discussion

Both general and specific features of the H. defensa genome reflect its lifestyle as a host-restricted, mutualist symbiont that invades host cells. The moderately reduced genome consists of a 2,110,331-bp circular chromosome and a 59,034-bp conjugal plasmid with average G + C contents of 40.1% and 45.3%, respectively (Table 1, Fig. 1). The chromosome contains a canonical origin of replication (oriC) situated between mmmG (gidA) and mioC. Of the 2,100 predicted coding sequences (CDS), 1,665 (79%) have homologs present in GenBank. Most remaining unique hypothetical proteins (75%) are <100 aa (AA), making their identity as true genes equivocal. In addition, 188 readily identifiable pseudogenes were present; this number is similar to that in Escherichia coli genomes (24).

Phylogenies based on single loci place H. defensa in the Enterobacteriaceae, but are otherwise poorly resolved (12, 25). In analyses of multigene alignments of conserved, single-copy, core proteins, H. defensa and another aphid endosymbiont, Regiella insecticola, consistently fell within a clade containing Yersinia spp. and Serratia spp. (Fig. 2). Low bootstrap values near these nodes are elevated by removing Hamiltonella and Regiella from analyses, suggesting that the long branches reduce confidence. Regardless, the phylogeny suggests that Hamiltonella and Regiella form a lineage distinct from the entomopathogenic nematode symbionts Photorhabdus and Xenorhabdus, and from the sequenced tsetse symbiont Sodalis glossinidius.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Data deposition: The sequences reported in this paper have been deposited in the GenBank database (accession nos. CP001277, CP001278).

1To whom correspondence should be addressed. E-mail: pdegnan@email.arizona.edu.

Author contributions: P.H.D., Y.Y., R.A.W., and N.A.M. designed research; P.H.D. and N.S. performed research; P.H.D. and R.A.W. analyzed data; and P.H.D. and N.A.M. wrote the paper.

This article contains supporting information online at www.pnas.org/cgi/content/full/0900194106/DCSupplemental.

www.pnas.org/cgi/doi/10.1073/pnas.0900194106

PNAS | June 2, 2009 | vol. 106 | no. 22 | 9063–9068
Table 1. Comparison of *H. defensa* genome features to those of relevant Enterobacteriaceae

<table>
<thead>
<tr>
<th></th>
<th><em>B. aphidica APS</em></th>
<th><em>H. defensa SAT</em></th>
<th><em>S. glossinidius</em></th>
<th><em>E. coli K12</em></th>
<th><em>Y. pestis CO92</em></th>
</tr>
</thead>
<tbody>
<tr>
<td>Chromosome, bp</td>
<td>640,681</td>
<td>2,110,331</td>
<td>4,171,146</td>
<td>4,639,221</td>
<td>4,653,728</td>
</tr>
<tr>
<td>Extrachromosomal elements</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>–</td>
<td>3</td>
</tr>
<tr>
<td>Total G + C (%)</td>
<td>26.2</td>
<td>40.1</td>
<td>54.7</td>
<td>50.8</td>
<td>47.6</td>
</tr>
<tr>
<td>Total predicted CDS</td>
<td>571</td>
<td>2,100</td>
<td>2,432</td>
<td>4,284</td>
<td>4,012</td>
</tr>
<tr>
<td>Coding density (%)</td>
<td>86.7</td>
<td>80.8</td>
<td>50.9</td>
<td>87.9</td>
<td>83.8</td>
</tr>
<tr>
<td>Average CDS size (bp)</td>
<td>984</td>
<td>812</td>
<td>873</td>
<td>950</td>
<td>998</td>
</tr>
<tr>
<td>Pseudogenes</td>
<td>13</td>
<td>188</td>
<td>972</td>
<td>150</td>
<td>149</td>
</tr>
<tr>
<td>rRNA operons</td>
<td>2</td>
<td>3</td>
<td>7</td>
<td>7</td>
<td>6</td>
</tr>
<tr>
<td>tRNAs</td>
<td>32</td>
<td>42</td>
<td>69</td>
<td>86</td>
<td>70</td>
</tr>
<tr>
<td>Lifestyle</td>
<td>Obligate</td>
<td>Facultative</td>
<td>Facultative</td>
<td>Commensal</td>
<td>Pathogen</td>
</tr>
</tbody>
</table>

CDS, coding sequences.

**Complementarity of Host and Symbiont Metabolisms.** The metabolism of *H. defensa* inferred from the genome confirms that it is host-dependent. It is an aerobic heterotroph that shares its central metabolic machinery with that of most free-living enteric bacteria (Fig. 3). Unlike most endosymbionts, *Hamiltonella* is also capable of the fermentation of pyruvate to lactate (pykF, *ldhA*) and acetyl-CoA to acetate (*pta, ackA*). Thus, *H. defensa* appears able to produce energy even under oxygen-limiting conditions.

Biosynthesis of essential amino acids and vitamins is a hallmark of nutritional endosymbionts, exemplified by *Buchnera*. Based on its gene set, *H. defensa* synthesizes only 2 essential and 7 nonessential amino acids, but can make most essential vitamins except thiamine (B1) and pantothenate (B5) (see Fig. 3). Unlike *Buchnera*, which lacks most active transport mechanisms, *H. defensa* likely acquires missing building blocks via substrate-specific transporters.

The essential amino acids that *H. defensa* requires are largely lacking from the insect diet of phloem sap (26). Our data suggest that both *H. defensa* and the host insect rely on *Buchnera*, the required endosymbionts that synthesize essential amino acids from this limited carbon and nitrogen source (9, 10). Except for the glutamate/aspartate transporter (*gltP*), the *H. defensa* genome contains no trace of the missing biosynthetic or transporter genes. This suggests that, unlike *S. glossinidius*, which very recently became host-restricted (27, 28), *H. defensa* has had a long-term association with insects (Table S1) consistent with previous evidence (12).

**Putative Virulence Mechanisms Involved in Symbiosis.** *H. defensa*’s abilities to invade novel insect hosts, to persist in them, and to kill their endoparasites are likely dependent on the presence of numerous loci commonly involved in pathogenicity (18). Our results give a complete inventory of these pathogenicity or symbiosis factors and indicate that some of these loci have been rearranged or disrupted. For example, *H. defensa* carries two
type-3 secretion systems (T3SS), which are similar in gene content and order to T3SS in Salmonella typhimurium LT2 (SPI-1, SPI-2) (18). These protein translocation systems are normally used by pathogens to invade host cells and evade host immune responses (29) and are required for the maintenance of the Sodalis-tsetse fly symbiosis (28, 30). Although both H. defensa T3SS are complete, neither forms a single genomic island. Putative secreted effector proteins are scattered throughout the genome and were probably acquired by multiple HGT events (Table S2).

The most abundant putative virulence factors are RTX (repeats in toxin) toxins: a protein family that includes a variety of exported proteins including α-hemolysin and leukotoxin (31). These proteins have highly variable lengths (800–6,000 AA) and contain a tandemly repeated nonapeptide sequence that is involved in binding calcium. The toxin genes (rtxA) tend to occur in operons containing an activating acyltransferase (rtxC) and an ABC transporter (rtxBD). H. defensa contains 32 CDS with similarity to rtxA, 2 copies of rtxB, and only a single copy of rtxD. These sequences are significantly diverged from known RTX toxins (20–40% AA identity), and several are possibly paralogs (60–92% AA identity). The rtxA copies include both intact (n = 10) and fragmented (n = 22) CDS. Together, these data suggest past duplication and diversification of these toxin genes, followed by mutation and inactivation of some copies.

Response of H. defensa to Changing Environments. Despite the constrained biosynthetic capabilities of H. defensa, it has considerably more cell structural, DNA replication, recombination, and repair genes than do obligate endosymbionts (2). H. defensa also retains more regulatory genes, including global regulators (e.g., 4 sigma factors), specific regulators of biosynthetic pathways (e.g., for production of biotin, cysteine, fatty acids), 4 pairs of putative 2-component regulators, and 3 genes involved in quorum sensing.

Pathogenic bacteria typically express virulence factors under strict regulatory controls. In H. defensa, putative regulatory genes flank both T3SS, one of which is homologous to hIL, the key regulator for SPI-1 (32). We have also identified homologs of Hha and SlyA, which activate the expression of hemolysins.

---

**Fig. 2.** Phylogenetic reconstruction of H. defensa and related Enterobacteriaceae using 88 single-copy orthologous proteins. Bacteria engaged in associations with insects are indicated (I). Support values are reported from 100 bootstrap replicates from RaxML and PhyML analyses values greater than 80 are indicated by asterisks.

---

**Fig. 3.** Metabolic reconstruction of H. defensa indicates that it can complete glycolysis, the tricarboxylic acid (TCA) cycle, and the pentose phosphate pathway, in addition to producing both pyrimidines and purines. Essential (red) and nonessential (green) amino acids are either synthesized de novo or imported by a substrate-specific transporter. Most vitamins and cofactors (blue) are synthesized, although pantothenate and thiamin must be imported. Circles indicate genes in a particular pathway that are present (filled) or absent (open). *Putative “polar” amino acid transporter may transport histidine or threonine.

---

Degnan et al.
(33, 34), and 2-component regulators and quorum-sensing genes are also known to influence expression of virulence factors. The diversity of regulatory genes suggests a mechanism by which *H. defensa* copes with changing environments, such as invasion of a new host species or attack of hosts by parasites.

**Repetitive Genomics.** The genome of *H. defensa* is riddled with mobile DNA. Insertion sequences (IS), group II introns, and plasmids comprise 21% of the genome (444,936 bp) (see Fig. 1). Estimates of genetic diversity for the integrated prophage, and plasmids comprise 21% of the genome mobile DNA. Insertion sequences (IS), group II introns, integration (see Table S3). The single active group II intron also suggests recent transpositional activity or gene conversion (see Table S3). Proliferation of repeats is expected in intracellular bacteria, as they tend to have small effective population sizes (*N_e*) because of recurrent transmission bottlenecks, increasing the level of genetic drift (35).

The lack of site specificity has resulted in retrotransposition within and between genes, as well as into previously retrotransposed group II introns. PCR screens of *H. defensa* strains from different hosts showed that ISHde1, ISHde2, and ISHde3 were widespread, whereas ISHde4 and the group II intron were in fewer than half of tested strains (see Table S3). Repetitations in *H. defensa*, as in many free-living bacteria, has been influenced by interactions with bacteriophage (23). Apart from the APSE prophage, *H. defensa* contains 22 phage-like gene blocks (153,384 bp), several of which have undergone partial duplication (see Fig. 1 and Table S4). The prophage islands were readily identified because of both gene content (e.g., phage integrases) and elevated G + C content (mean 46.5%). Except for APSE, the prophage appear to be inactive, as all of the islands are fragmentary and most contain inactivated or truncated genes. Mobile elements were probably involved in the inactivation, rearrangement, and duplication of the gene blocks, most of which (16 of 22) are flanked on one or more sides by either an IS element or group II intron. *H. defensa* bears a conjugative IncFII plasmid pHD5AT. The type I secretion system (T4SS) and pilus it encodes is similar to the *tra* and *pil* loci from the *Serratia entomophila* plasmid pADAP. The loci in *H. defensa* bear a number of hypothetical genes (Fig. 4). In contrast to pADAP, the pHD5AT plasmid has no genes implicated in virulence or resistance.

**Integrated plasmid genes represent 9% of the *H. defensa* genome (197,022 bp) (see Fig. 1 and Table S4). They share features with the prophage blocks, including elevated G + C content (44.2%), a large fraction of pseudogenes or truncated proteins, and flanking IS elements or group II introns. Two of the islands are the result of chromosomal integration and decay of pHD5AT, as indicated by missing or inactivated genes (Fig. S1 and SI Methods). Two other plasmid islands are inactivated T4SS, yet are phylogenetically distinct from the *tra* locus on pHD5AT (see Table S4). The remaining islands contain a variety of plasmid-associated genes, but precise assignation of fragments to plasmids or integration events are difficult because of recombination.

**H. defensa Proteome.** To explore the expression of *H. defensa* genes and proteins, we performed a proteomics experiment on a sample of purified *H. defensa* cells, using the genome sequence for peptide and protein identification. Implementing conservative identity cutoffs, we identified 89 expressed proteins (Fig. 4 and Table S5). Several phage APSE proteins and one TSS protein (SscE) were recovered. Among the most highly expressed proteins were those involved bacterial responses to stress and membrane components. Indeed, the most abundant protein, GroEL (Hisp60, MopA) or chaperonin, is also the most abundant protein in other obligate and facultative endosymbionts (37).

Other recovered *H. defensa* proteins include ones involved in core processes (e.g., transcription, translation) and conserved or hypothetical proteins encoded in the genome but having unknown functions.

**Conclusions**

The reduced size and compositional bias in the genome of *H. defensa* reflects a long-term, stable association with its insect hosts. In this respect, the *H. defensa* genome is similar to *Wolbachia* genomes, which are small, have highly reduced biosynthetic capabilities, and encode an abundance of mobile genetic elements (3, 5). Whereas *Wolbachia* is known mostly as a reproductive parasite and antagonist of its hosts, *H. defensa* protects hosts from parasites. Genes for toxins, effector proteins, and TSS are likely to be critical elements underlying this mutualistic role. The presence of numerous homologs of known virulence factors, which have homologs in other insect symbionts and in mammalian and plant pathogens, reiterates how conserved genetic mechanisms involved in bacterial-eukaryotic cellular interactions can result in vastly different outcomes. Some of the virulence-gene homologs (e.g., *traA*) are not intact, suggesting a changing role for the toxins in this symbiosis. These shifting gene sets likely reflect the inherent dynamism of antagonistic interactions, which impose ongoing selection for counter-adaptations in parasites, hosts, and symbionts. Gene losses and inactivations in *H. defensa* are tempered by gene gains via HGT, evidenced by the abundance of plasmid and phage islands. Although the variable toxins encoded by the phage APSE appear to contribute to parasitoid protection, the *H. defensa* genome reveals a history of association with other phage and plasmids that likely played an earlier role in resorting ecologically important genes among *H. defensa* strains and possibly other bacteria.

**Methods**

**DNA Isolation and Construction of Libraries.** We used 2 complementary sequencing strategies to complete the *H. defensa* genome: (i) subcloning and Sanger sequencing a large insert BAC library and (ii) pyrosequencing (Fig. S2). Intact *H. defensa* cells were purified from whole insects to minimize contamination with aphid and *Buchnera* DNA, as described previously (18). A BAC library was constructed, fingerprinted, and minimal tiling paths were chosen (as in ref. 38). Individual BACs were then subcloned, sequenced bidirectionally with ABI3730xl sequencers, and assembled using Phred, Phrap, and Consed (39–41). Overlapping and validated BACs were then merged.

**Bacterial genomes contain nonclonable fragments, so we performed pyrosequencing as an unbiased sequencing method. High molecular weight DNA was isolated directly from the purified *H. defensa* cells using the Puregene Tissue Core Kit B (Qiagen). We generated a standard and paired-end single-stranded template DNA (stdDNA) library using the GS DNA Library
Preparation Kits (Roche Applied Sciences) that were then amplified by emPCR and sequenced on a GS-FLX (454 Life Sciences). The 454 reads were assembled with Newbler (v1.1.0.3.24) using default parameters.

Final Assembly and Genome Closure. Putative H. defensa contigs generated from the 454 reads and distinct from the finished BACs were sorted and oriented using linking information from the paired ends. PCR primers were designed at the contig ends, and products were amplified and sequenced using standard protocols described elsewhere (23). The 454 reads for each scaffold were then reassembled with Newbler, and Sanger reads were incorporated in Consed using Phrap.

Genome Annotation. Genes were predicted for the finished H. defensa genome using Glimmer v3.02 (protein-coding genes), TRNAscan-SE (tRNAs) and Blastn (structural and ribosomal RNAs). Putative CDS greater than 30 AA were annotated using consensus of BlastP similarity searches to NR, all microbial genomes, and E. coli str. K12 and protein domain searches using Hmmr and the Pfam.Is database (42). CDS without hits having expectation values less than $10^{-10}$ (BlastP) and $10^{-4}$ (Pfam) were annotated as hypothetical, and CDS with conflicting results were assigned as putative. Predicted start codons were adjusted manually using alignments to the top 5 NR hits and the E. coli best hit if present. Intergenic regions were rescreened with BlastX for possible CDS missed by Glimmer. CDS with truncations > 40% length or fragmented CDS were designated pseudogenes in the final annotation. Boundaries of multi-copy repeats (e.g., insertion sequences, group II introns) were identified by consensus alignments. Gene functions were inferred from those of identified homologs, and the integration of genes into metabolic pathways was determined using EcoCyc (43).

Whole Genome Phylogeny. Multigene phylogenetic reconstruction was used to determine the relationship of H. defensa with other gamma-proteobacteria. Briefly, we identified 88 of 203 single copy orthologs (SICO) in H. defensa and 29 other genomes (Table S6) (44). Protein sequences of each orthology were aligned in Muscle v3.6.45, and all invariant and gap-containing columns were removed. Individual protein alignments were then concatenated into 4 alignments (Table S7). Alignments without H. defensa and R. insecticola sequences were also generated to assess impact of of long-branch attraction or other artifacts. Each dataset was analyzed with RaXML and PhyML (46, 47), and unique topologies were compared using the SH-test in TREE-PUZZLE 5.2 (48). The topology with the lowest log likelihood and that disagreed with the fewest datasets is presented. Support values were estimated from 100 non-parametric bootstrap replicates.

Protein Expression. Briefly, H. defensa cells were isolated as above and immediately frozen at ~80°C. The cell pellet was thawed, homogenized, and centrifuged, and proteins were precipitated. The resulting pellet was dissolved and run on a 10% SDS/PAGE gel, and the lane was divided into sections and transferred to a Nitrocellulose membrane by vacuum transference. The tephritid lytic predation. The tephritid Tephritis brassicae was extracted from each gel section, concentrated, and injected into an LC/MS/MS system. Resultant tandem mass spectra were processed and analyzed with Mascot 2.2 (Matrix Science), using a database of H. defensa, B. aphidicola, and A. pismum protein sequences. The results were filtered using a Mascot significance threshold of 0.05 and Mowse ion score cutoff of >31, and the false-discovery rate for H. defensa peptides was 0.2%.

ACKNOWLEDGMENTS. The authors thank K. Hammond, B. Nankivell, K. Sunitsch and J. Currie, T.R. Mueller, K. Collura, R. He, and J.L. Goicoechea of the Arizona Genomics Institute. We also thank J. Evbank and G. Goodrich-Blair for access to unpublished genome data and Q. Lin at the University of Albany Proteomics Facility for running the protein sample. This research was supported by National Science Foundation Grant 0313737 (to N.A.M.). P.H.D. received funding from National Science Foundation Integrative Graduate Education and Research Traineeship Fellowship in Evolutionary and Functional Genomics, the Center for Insect Science at the University of Arizona, and National Science Foundation Doctoral Dissertation Improvement Grant Award 0709992.

Degnan et al.