27. J. H. McDonald, M. Kreitman, *Nature* **351**, 652 (1991).
28. F. Tajima, *Genetics* **123**, 585 (1989).
29. We thank C. Aquadro, J. Fry, R. Glor, H. Malik, D. Presgraves, Y. Tao, and the Orr lab for discussions; D. Lambert for advice on molecular techniques; H. Malik and N. Elde for help with PAML; and S. Schaeffer for fly stocks. Financial support was provided by NIH (GM51932). All sequences are available at GenBank (accession numbers FJ349335 to FJ349342 and FJ418600 to FJ418631). The authors declare no conflict of interest.

Supporting Online Material
www.sciencemag.org/cgi/content/full/1163934/DC1
Materials and Methods
SOM Text
Figs. S1 to S7
Tables S1 to S9
References

29 July 2008; accepted 7 November 2008
Published online 11 December 2008;
10.1126/science.1163934
Include this information when citing this paper.

The Dynamics and Time Scale of Ongoing Genomic Erosion in Symbiotic Bacteria

Nancy A. Moran,[1]* Heather J. McLaughlin,[1] Rotem Sorek[2]

Among cellular organisms, symbiotic bacteria provide the extreme examples of genome degradation and reduction. However, only isolated snapshots of eroding symbiont genomes have previously been available. We documented the dynamics of symbiont genome evolution by sequencing seven strains of *Buchnera aphidicola* from pea aphid hosts. We estimated a spontaneous mutation rate of at least 4×10^{-9} substitutions per site per replication, which is more than 10 times as high as the rates previously estimated for any bacteria. We observed a high rate of small insertions and deletions associated with abundant DNA homopolymers, and occasional larger deletions. Although purifying selection eliminates many mutations, some persist, resulting in ongoing loss of genes and DNA from this already tiny genome. Our results provide a general model for the stepwise process leading to genome reduction.

Obligate symbionts and pathogens, which have evolved repeatedly from free-living bacterial ancestors, show striking convergence in fundamental genomic features. In several symbionts of insects, most ancestral genes are eliminated by deletion, resulting in some of the smallest known cellular genomes (*1*–*4*). Symbionts also display rapid evolution at both the DNA and peptide sequence levels and have highly biased nucleotide base compositions, with elevated frequencies of adenine and thymine (A+T). Because these genomes are asexual and do not acquire foreign DNA, each gene loss is irreversible (*2*, *5*–*7*). These genomic features have been ascribed to increases in genetic drift associated with a host-restricted life-style (*8*, *9*) and, potentially, to an increased mutation rate and biased mutational profile stemming from the loss of DNA-repair genes, which are among the gene categories most depleted in symbiont genomes (*1*, *10*).

Although numerous sequenced examples of reduced genomes in obligate symbionts or pathogens are available, these are too distantly related to permit stepwise reconstruction of genomic changes. As a result, the dynamics of ongoing genomic erosion, the extent to which mutation rate is elevated, the effectiveness of natural selection in purging mutations, and the nature of the mutational events that lead to further loss of DNA and metabolic functions are unclear. To illuminate these evolutionary processes, we sequenced several closely related genomes of the obligate symbiont *Buchnera aphidicola* from a single host species, the pea aphid *Acyrthosiphon pisum* (*Buchnera-Ap*). A previously sequenced genome of *Buchnera-Ap* showed a gene set typical for an obligate symbiont (*1*) lacking most ancestral genes, including genes underlying transcriptional regulation, biosynthesis of cofactors present in hosts, DNA repair, and other processes. The 607 retained genes encode machinery for replication, transcription, translation, and other essential processes, as well as biosynthetic pathways for essential amino acids required by hosts (*1*).

A. pisum is native to Eurasia, but has been introduced worldwide. It was first detected in North America in the 1870s (*11*). We sequenced the genomes of seven *Buchnera-Ap* strains descended from two colonizers of North America (and hence diverging up to 135 years ago), including two strains diverging in the laboratory for 7.5 years. Solexa sequencing was combined with verification by Sanger sequencing (*12*), to determine genomic sequences of these seven strains (Table 1). A total of 2392 positions (0.3% of sites on the 641-kb chromosome) showed a nucleotide substitution. These single-nucleotide polymorphisms (SNPs) were distributed approximately evenly around the chromosome (fig. S1). We also detected a total of 149 insertion or deletion events (indels): 134

[1]Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, AZ 85721, USA. [2]Department of Molecular Genetics, Weizmann Institute of Science, Rehovot 76100, Israel.

*To whom correspondence should be addressed. E-mail: nmoran@email.arizona.edu

Table 1. Description of sequence data.

	Tuc7	9-2-1	8-10-1	8-10-1	A2A	5AR	5A	7A
Source locality, year	Tucson AZ, 1999	Cayuga Co NY, 2001	Cayuga Co NY, 2001	Cayuga Co NY, 2001	Logan UT, 2003	Derived from 5A, 2000	Madison WI, 1999	Cayuga Co NY, 2000
Average read size	39	36	39	36	39	36	39	39
No. of initial reads	3,185,491	11,293,714	9,064,851	20,653,949	9,224,134	9,615,693	8,227,047	18,234,517
No. of reads mapped	1,024,330	6,731,726	4,432,760	12,977,253	7,088,978	6,944,135	4,448,342	12,150,323
Percent reads mapped*	32.16	59.61	48.90	62.83	76.85	72.22	54.07	66.63
Fold genome coverage (average)	57.9	369.9	259.4	700.5	412.2	374.6	256.0	661.1

*Unmapped reads represent contaminating DNA, largely from the host genome.

single-base indels, 12 indels of 2 to 16 bases, and 3 large deletions (220 to 1131 bases), also dispersed around the genome (fig. S1).

Parsimony analysis of SNPs yielded a single phylogenetic tree with no homoplasy, as expected for clonal lineages if each base substitution is a singular event (Fig. 1A). Indels showed almost no homoplasy; all but two mapped as single events (Fig. 1B). The newly sequenced genomes comprised two tight clusters that were divergent from each other and even more divergent from the reference strain, Tokyo1998 (Fig. 1A). Rooting the phylogeny on the branch leading to Tokyo1998 enabled us to assign direction of change for both base changes and indels on the lineages leading to the two clusters.

We inferred that the two clusters consist of descendants of two separate female colonizers, each arriving in North America sometime after 1870, by constructing a phylogeny on the basis of a 1.1-kb DNA fragment from 38 *Buchnera-Ap* samples collected in America, Asia, and Europe (*12*). The clades corresponding to these two clusters contain the large majority of North American samples, but are absent (Cluster 2) or rare (Cluster 1) among samples from Eurasia, where diverse lineages are present (fig. S2). Thus, we dated the common ancestor of each cluster representing an introduced matriline to a maximum of 135 years ago. Averaging the pairwise divergences through the ancestral node of each cluster, we calculated rates of nucleotide substitution of 19 SNPs per genome per 270 years for each U.S. haplotype cluster (divergence times are doubled to derive the rate along a single evolving lineage). After pooling the observed changes, we estimated the rate as 0.70 substitutions per genome per decade [95% confidence interval (CI): 0.51 to 0.97 substitutions per genome per decade], or 1.1×10^{-7} substitutions per site per year (95% CI: 0.8 to 1.5×10^{-7} substitutions per site per year). The rate is doubled (2.2×10^{-7}; 95% CI: 1.4 to 3.3×10^{-7} substitutions per site per year) if calculated on the basis of changes at intergenic spacers and synonymous sites, that is, genomic sites that generally can tolerate mutations with little effect on fitness and that are thus expected to approximate the mutation rate (Table 2) (*12*). Adjusting for *Buchnera* replications per year [by estimating *Buchnera* divisions per aphid generation and aphid generations per year (*13*)] gives an estimate of 4×10^{-9} substitutions per site per replication.

Our estimated mutation rate for base changes was unexpectedly high: more than 10 times the previous estimates of mutation rate calculated on the basis of silent site divergences in both *Buchnera* and free-living bacteria (*5*, *13*). Although several artifacts could affect these calculations, the main source of error, a more recent coalescence of introduced clusters than estimated, would actually make this an underestimate of the mutation rate. Also, even spacers and silent sites may be subject to some purifying selection. Thus, the rate of spontaneous mutation (or substitution at neutral sites) is almost certainly higher than our estimates. A high mutation rate was also supported by the finding of two base substitutions fixed in a laboratory line (5AR) during 7.5 years (Fig. 1A). Although previous estimates of mutation rate in *Buchnera*, from genome pairs diverging 60 million years ago, were lower, those calculations were unreliable because intergenic spacers were too divergent to allow alignment and because silent sites underwent too many substitutions for accurate estimation of divergence (*5*).

This rate calibration can be used to estimate divergence times of older lineages of *Buchnera-Ap* used in this study. If the root of the tree is on the branch leading to the Tokyo1998 strain, we calculated that the lineage leading to the two clusters we sequenced (showing an average divergence of 1617 substitutions per genome) diverged from the Tokyo1998 strain 11,489 (95% CI: 8340 to 15,790) years ago. Calculations made only on the basis of intergenic spacers and synonymous changes gave similar estimates [12,555 (95% CI: 8292 to 20,030) years].

We next considered trends in nucleotide-composition bias. A distinctive feature of most small bacterial genomes is an elevated A+T content, reflecting biased mutational patterns. Our data show no evidence of continued evolution toward increased A+T content in *Buchnera-Ap*. The 50 substitutions in the terminal branches of the tree had little effect on base composition, with 21 increasing, 31 decreasing, and 6 not affecting A+T content (Fig. 1A). For the 1423 substitutions on the branches leading to the two clusters, base composition was in near-equilibrium, with 47% decreasing and 48% increasing overall G+C content. This implies that the overall genomic base composition near 25% G+C is an approximate equilibrium, consistent with a mutation rate from G/C to A/T that is three times as high as that for A/T to G/C (Fig. 2C) (*14*). However, this equilibrium could be disturbed if additional DNA repair functions are lost.

To estimate the effect of purifying selection on the ongoing evolution of *Buchnera-Ap* genomes, we analyzed base substitutions in coding regions. On the basis of the genome-wide frequencies of mutation types acting on each base (Fig. 2C) and codon frequencies calculated for the entire *Buchnera-Ap* genome, mutations causing amino acid replacements are expected to arise 4.5 times as often as those affecting only codon choice (*12*). But only 36% of observed substitutions were replacement substitutions, giving a per-site ratio of replacement to silent changes (dN/dS) of 0.125 and implying that most mutations affecting polypeptide sequence are purged by selection. Purifying selection was also evident from the concentration of indels in intergenic spacers, which are largely selectively neutral regions sometimes recognizable as eroding pseudogenes (*5*). Of 146 small indels, 134 (92%) occur in intergenic spacers, which constitute only 13.5% of the genome. The SNP-to-indel ratio is 3.1 in spacers and 166.4 in coding regions (Table 2). This paucity of indels within coding genes reflects the fact that most indels cause frameshifts, leading to dysfunctional protein products, and are eliminated by selection. Indeed, 11 of 12 indels observed in coding regions imposed frameshifts. Thus, during the evolution of these *Buchnera-Ap* lineages, an estimated 82% of new indels have been purged by selection.

To determine whether genome erosion is ongoing in the closely related genomes that we sequenced, we addressed whether the 146 detected indels contributed to genome reduction. Our findings indicate that small indels do not directly cause DNA loss in *Buchnera*, because

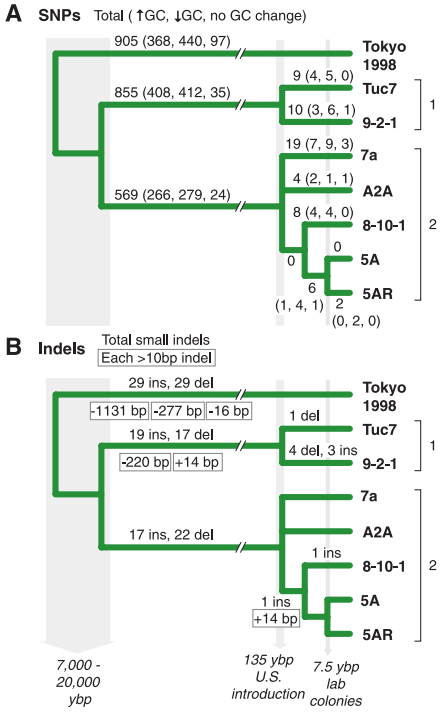


Fig. 1. A maximum-parsimony tree showing the evolutionary reconstruction of phylogenetic relationships and changes in the genomes of *Buchnera-Ap* strains. (**A**) Single-nucleotide substitutions. (**B**) Small and large insertion and deletions.

Table 2. Mutational patterns in genomes of *Buchnera* symbionts of pea aphids, from all base changes and insertion/deletion events, of Fig. 1.

	SNPs	SNPs/kb	Small indels	Indels/kb	SNP/indel ratio
Protein-coding	1997	3.6	12	0.02	166.4
Intergenic spacer	395	4.6	130	1.50	3.1
Total	2392	3.7	142	0.22	16.8

small deletions were balanced by small insertions on all branches of the tree, regardless of root position (Fig. 1B). But three large deletions did effect a net DNA loss: (i) a 220–base pair (bp) deletion in the *znuC-pykA* spacer in the lineage leading to Cluster 1; (ii) a 277-bp deletion in the *gapA-fidA* spacer in the lineage leading to Tokyo1998; and (iii) a 1131-bp deletion corresponding to part of the gene *yaeT* and the entire sequence of the gene *fabZ*, also in the lineage leading to Tokyo1998. All large deletions corresponded to positions of extra genes (*znuA-yebA*, *queF*, and *fabZ*) in the *Buchnera–Schizaphis graminum* genome (*5*), suggesting that the detected deletions eliminated relics of these genes from the *Buchnera-Ap* genomes. No large insertions were identified, consistent with previous evidence that *Buchnera* does not acquire foreign genes (*2*, *5*, *6*). Together these three deletions account for a loss of 1625 nucleotides (Fig. 1B). This corresponds to DNA loss at a rate of roughly 1 kb per 10,000 years, although this estimate is subject to a high error rate due to the small number of large deletions observed.

We next studied whether ongoing loss of functional genes has occurred during the divergence of the *Buchnera-Ap* genomes. We observed 16 genes that appear to be inactivated, either through a 1- to 2-base indel causing a frameshift (11 genes), a base substitution generating a stop codon (three genes), or a large deletion (one event, two genes) (table S1). Functions of these genes include DNA repair (*ung*, *sbcB*), biosynthesis potentially affecting host amino acid or vitamin nutrition (*argC*, *trpB*, *glyA*, *ribD2*), fatty acid biosynthesis (*fabZ*), or cell envelope production (*murF*), and genes involved in transport (*ynfM*) or secretion (*fliK*, *flgB*) (table S1). Genes in these functional categories have been noted to undergo degradation or loss in distantly related strains of *Buchnera* and other obligate symbionts (*2*, *5*, *6*). Such losses could influence further genomic evolution (by affecting mutation), as well as the ability to provision hosts. Genes with frameshift mutations may retain partial functionality, through production of some in-frame transcripts due to slippage of RNA polymerase (*15*), but the notable concentration of indels in intergenic spacers implies that most frameshifts adversely affect gene function.

As indicated above, SNPs are about three times as common among new mutations as are small indels (Table 2), but indels mediated most inferred gene inactivations. Small indels were heavily concentrated in mononucleotide runs ("homopolymers"), with 93% of single-base indels linked to runs of at least five and 66% in runs of at least seven consecutive A's or T's (Fig. 2, A and B). [Solexa sequencing resolves homopolymer length with low error and without bias (*12*).] The incidence of indels per run was highest in the longest runs (10 and above), but because the longest runs were rare, most indels were found in runs of 7 to 9 bases (Fig. 2A, B).

Together, the evolutionary trends observed in *Buchnera-Ap* converged to a model describing a stepwise process of symbiont genome erosion (Fig. 3). The shift toward high A+T content that is common in host-restricted bacteria leads to increased occurrence of A/T homopolymers. These, in turn, are hot-spots for small indels, which are elevated in homopolymers due to replication slippage, and which are further increased when certain DNA repair pathways are compromised (*16*). Our data imply that many new indels disrupt reading frames and that most are removed by selection. However, a minority persists, leading to inactivated genes. The resulting pseudogenes undergo rapid sequence evolution due to the lack of purifying selection and are eventually removed by large deletions. Because large deletions do not precisely excise inactivated genes, intergenic spacers often persist in the positions of former genes.

This process fits well with previous observations comparing more distantly related symbiont genomes [e.g., (*5*–*7*)]; however, those studies lacked the precision needed to detect the critical role of homopolymers and frameshifts in gene inactivation. Our model predicts that the initial step leading to genome reduction is a shift in nucleotide composition toward higher A+T content. Loss of DNA-repair functions has been proposed as the cause for this shift (*17*). A consequence of high A+T content is an excess of homopolymers and a resulting high incidence of small indels (Fig. 2, A and B) leading to gene inactivations. Indeed, A+T-biased genomes, including *Buchnera* genomes, show higher frequencies of A/T homopolymers than expected by chance alone (*18*), reflecting mutational patterns that yield longer A/T runs through replication slippage or other processes (*16*).

Most sequenced insect symbiont genomes are between 0.6 and 1 megabases in size and contain more than 500 genes, similar to the smallest known pathogen genomes and consistent with previous suggestions that cellular genomes have a minimal size threshold (*1*, *5*, *6*, *19*, *20*). This study, as well as the recent discovery of symbiont genomes containing only 182 to 450 genes (*2*–*4*), suggests instead that the process of gene loss has no clearly defined limit. We identified a surprisingly high rate of new mutations, including both base changes and indels, in the genomes of *Buchnera-Ap*. Although most mutations impairing gene function are removed by selection, others persist, leading to the permanent inactivation of genes and the subsequent loss of the corresponding DNA through larger deletions.

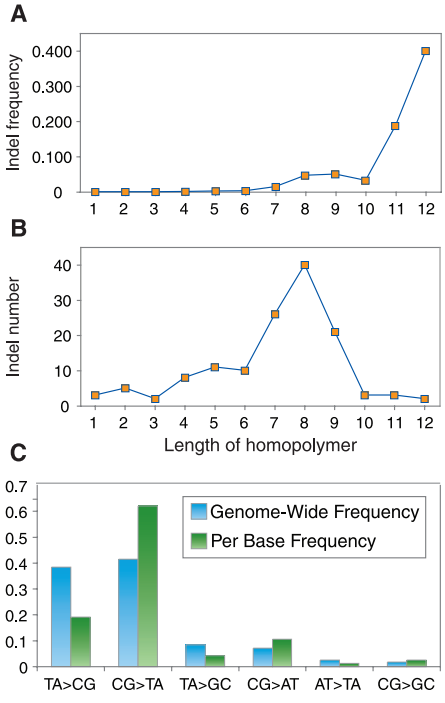


Fig. 2. Observed mutations in *Buchnera-Ap*. Frequencies of single-base indels in homopolymers of different lengths (**A**) per homopolymer and (**B**) genome-wide. (**C**) Relative frequencies of base substitutions genome-wide and per nucleotide base.

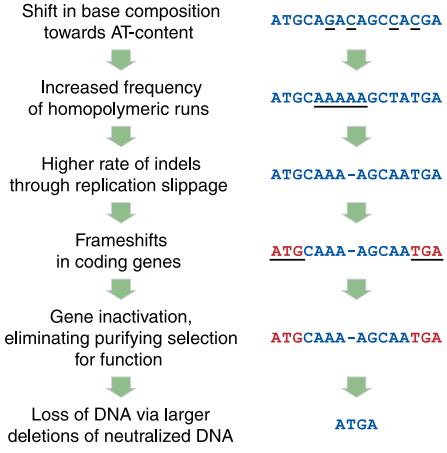


Fig. 3. Model of symbiont genome erosion, from mutational patterns revealed by sequencing the complete genomes of seven *Buchnera-Ap* strains.

References and Notes

1. S. Shigenobu, H. Watanabe, M. Hattori, Y. Sakaki, H. Ishikawa, *Nature* **407**, 81 (2000).
2. V. Perez-Brocal *et al.*, *Science* **314**, 312 (2006).
3. J. P. McCutcheon, N. A. Moran, *Proc. Natl. Acad. Sci. U.S.A.* **104**, 19392 (2007).
4. A. Nakabachi *et al.*, *Science* **314**, 267 (2006).
5. I. Tamas *et al.*, *Science* **296**, 2376 (2002).
6. R. C. H. J. van Ham *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **100**, 581 (2003).
7. P. H. Degnan, A. B. Lazarus, J. J. Wernegreen, *Genome Res.* **15**, 1023 (2005).
8. N. A. Moran, *Proc. Natl. Acad. Sci. U.S.A.* **93**, 2873 (1996).
9. T. Hosokawa *et al.*, *PLoS Biol.* **4**, e337 (2006).
10. N. A. Moran, J. P. McCutcheon, A. Nakabachi, *Annu. Rev. Genet.* **42**, 165 (2008).
11. A. M. Harper, J. P. Miska, G. R. Manglitz, B. J. Irwin, E. J. Armbrust, in *Special Publication 50* (Agricultural Experimental Station, University of Illinois, Urbana-Champaign, 1978), pp. 1–89.
12. Materials and methods are available as supporting material on *Science* Online.
13. H. Ochman, S. Elwyn, N. A. Moran, *Proc. Natl. Acad. Sci. U.S.A.* **96**, 12638 (1999).
14. N. Sueoka, *Proc. Natl. Acad. Sci. U.S.A.* **85**, 2653 (1988).
15. I. Tamas *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **105**, 14934 (2008).

16. B. O. Parker, M. G. Marinus, *Proc. Natl. Acad. Sci. U.S.A.* **89**, 1730 (1992).
17. N. A. Moran, *Cell* **108**, 583 (2002).
18. T. Coenye, P. Vandamme, *DNA Res.* **12**, 221 (2005).
19. L. Akman *et al.*, *Nat. Genet.* **32**, 402 (2002).
20. D. Wu *et al.*, *PLoS Biol.* **4**, e188 (2006).
21. We thank staff at the Joint Genome Institute for the Solexa sequencing runs, E. Rubin for support and facilities, and H. Ochman for comments on the manuscript. Aphid samples were donated by G. Powell and J. C. Simon. K. Hammond maintained insect cultures, and B. Nankivell prepared figures. Funding was provided by NSF award 0723472 (to N.A.M.), the Y. Leon Benoziyo Institute for Molecular Medicine (to R.S.), and the U.S. Department of Energy (to the Joint Genome Institute supporting the sequencing). GenBank accession numbers are CP00158 and CP00161 for the genomic sequences and FJ357501 to FJ357538 for sequences used for the phylogeny of fig. S2.

Supporting Online Material
www.sciencemag.org/cgi/content/full/323/5912/379/DC1
Materials and Methods
Figs. S1 and S2
Table S1
References

13 October 2008; accepted 19 November 2008
10.1126/science.1167140

A Polymorphism in *npr-1* Is a Behavioral Determinant of Pathogen Susceptibility in *C. elegans*

Kirthi C. Reddy,[1]* Erik C. Andersen,[2,3]* Leonid Kruglyak,[2,3]† Dennis H. Kim[1]†

The nematode *Caenorhabditis elegans* responds to pathogenic bacteria with conserved innate immune responses and pathogen avoidance behaviors. We investigated natural variation in *C. elegans* resistance to pathogen infection. With the use of quantitative genetic analysis, we determined that the pathogen susceptibility difference between the laboratory wild-type strain N2 and the wild isolate CB4856 is caused by a polymorphism in the *npr-1* gene, which encodes a homolog of the mammalian neuropeptide Y receptor. We show that the mechanism of NPR-1–mediated pathogen resistance is through oxygen-dependent behavioral avoidance rather than direct regulation of innate immunity. For *C. elegans*, bacteria represent food but also a potential source of infection. Our data underscore the importance of behavioral responses to oxygen levels in finding an optimal balance between these potentially conflicting cues.

Microbes, including commensal organisms and pathogens, profoundly influence the immune and metabolic physiology of host organisms (*1*). We used the nematode *Caenorhabditis elegans* as an experimental host to dissect the molecular basis of interactions between host species and microorganisms. *C. elegans* exhibits diverse behaviors in response to bacteria provided as a nutrient source (*2*–*4*). Feeding behavior can be modulated by environmental conditions, including oxygen concentration (*5*). Some bacterial species are pathogenic to *C. elegans* (*6*), and *C. elegans* responds by activating conserved innate immune pathways (*7*–*9*) and avoiding pathogens (*10*–*12*).

We found that the standard laboratory strain N2 (isolated in Bristol, England) and strain CB4856 (isolated in Hawaii, USA) exhibited a marked difference in susceptibility to the human opportunistic pathogen *Pseudomonas aeruginosa* strain PA14 (Fig. 1A) (*13*). The mean time to 50% lethality (LT50) for CB4856 was shorter (50 ± 7.8 hours) than that for N2 (90 ± 13 hours). Using a collection of recombinant inbred lines (*14*), we mapped the pathogen susceptibility trait to a 774-kb region of the X chromosome (LGX) containing *npr-1*, which encodes a G protein–coupled receptor related to the mammalian neuropeptide Y receptor (Fig. 1B). The 215V *npr-1* allele in N2 has increased NPR-1 activity relative to the 215F *npr-1* allele in CB4856, and the 215V allele confers behavioral differences that are dominant to those conferred by the 215F allele (*15*). To test the possibility that *npr-1* causes the difference in pathogen susceptibility between the N2 and CB4856 strains, we used *npr-1* loss-of-function mutants isolated in the N2 background. Like the CB4856 strain, the *npr-1* presumptive null alleles *ad609* and *ky13*, along with the reduction-of-function alleles *ur89* and *n1353*, had enhanced susceptibility to killing by PA14 (Fig. 1C and fig. S1). The enhanced susceptibility of *npr-1(ky13)* mutants was rescued by a transgene containing N2 wild-type (WT) copies of *npr-1*, and an N2 *npr-1* null mutation failed to complement the pathogen susceptibility phenotype of CB4856 (Fig. 1C). Thus, the enhanced susceptibility to pathogen of CB4856 is caused by the ancestral 215F allele of *npr-1*. This finding is consistent with a recent report by Styer *et al.* (*16*) that showed that loss-of-function mutations in the *npr-1* gene in the N2 background result in enhanced susceptibility to pathogen killing.

The 215F *npr-1* allele in CB4856 and loss-of-function mutations in *npr-1* confer a constellation of related behavioral phenotypes that have been termed "social feeding"—the animals associate together in groups (clumping) and are often found at the edge of the bacterial lawn (bordering) (*15*). The characterization of aerotaxis behavior in *C. elegans* revealed that CB4856 and *npr-1* loss-of-function mutants prefer the decreased oxygen concentrations found at the edge of the live bacterial lawn, which drives the bordering phenotype (*5*, *17*, *18*). We hypothesized that differences in behavior, instead of in innate immune responses as recently proposed (*16*), might underlie the observed pathogen susceptibility differences caused by the *npr-1* polymorphism. By spending more time on the bacterial lawn, CB4856 and *npr-1* mutants would receive an increased dose of the pathogenic bacteria, leading to higher mortality. Multiple independent experiments support our hypothesis.

First, mutations in the oxygen-sensing guanylate cyclase *gcy-35* and the neuronal signaling genes *ocr-2* and *osm-9*, which are necessary for *npr-1*–mediated bordering and aerotaxis behaviors (*5*, *19*, *20*), also suppressed the pathogen susceptibility of *npr-1* mutants (fig. S2, A and B). These data suggest that the clumping and bordering behaviors mediated by responses to oxygen concentration are necessary for the enhanced susceptibility to the pathogen.

Second, we altered the standard slow-killing pathogenesis assay (*21*) by spreading the PA14 lawn to the edges of the agar plate. In this "big lawn" assay, there is no region of the plate in which the animals can avoid pathogen. Under these conditions, N2 displayed increased susceptibility equivalent to both CB4856 and *npr-1(ad609)*, whereas the susceptibilities of *npr-1(ad609)* and *npr-1(ky13)* were equivalent in both assays (Fig. 2A and figs. S3 and S4). These data demonstrate that the pathogen susceptibility difference arises not from differential activation of immune pathways, but rather from the aberrant aerotaxis behavior of the N2 strain in the presence of the pathogenic lawn that results in lower exposure to the pathogen.

Third, we carried out the standard pathogenesis assay at 10% oxygen concentration, which suppresses bordering and aerotaxis behaviors in CB4856 and *npr-1* mutants (*5*, *18*). We observed that this reduced oxygen concentration also suppressed the pathogen susceptibility phenotypes of CB4856 and *npr-1(ad609)*. Allowing CB4856 and *npr-1* mutants to disperse off of the bacterial lawn results in survival that is equivalent to that observed for N2.

In addition, we found that three dauer-defective mutants that weakly aggregate and bor-

[1]Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. [2]Howard Hughes Medical Institute, Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, NJ 08544, USA. [3]Department of Ecology and Evolutionary Biology, Princeton University, Princeton, NJ 08544, USA.

*These authors contributed equally to this work.
†To whom correspondence should be addressed. E-mail: leonid@genomics.princeton.edu (L.K.); dhkim@mit.edu (D.H.K.)

Supporting Online Material for

The Dynamics and Time Scale of Ongoing Genomic Erosion in Symbiotic Bacteria

Nancy A. Moran,* Heather J. McLaughlin, Rotem Sorek

*To whom correspondence should be addressed. E-mail: nmoran@email.arizona.edu

Published 16 January 2009, *Science* **323**, 379 (2009)
DOI: 10.1126/science.1167140

This PDF file includes:

Supporting online materials

Materials and Methods

Each aphid strain consisted of a parthenogenetic clone descended from a single field-collected female and lab-grown at constant 20°C on fava bean (*Vicia faba*) seedlings. Localities and dates of strain collections are listed in Table 1. All strains were collected from alfalfa, *Medicago sativa*, except for 5A, which was from bur-clover, *Medicago lupulina*.

Like other obligate symbionts, *Buchnera* has not been cultured independently from hosts, so resequencing was performed with DNA from aphid-symbiont genomic mixtures. Before DNA extraction, enrichment for *Buchnera* cells was achieved with Percoll gradients and filtration, as described (S1). A total of about 5 ug DNA was obtained from each sample, starting with approximately 3 g of aphids.

Solexa/Illumina sequencing technology (S2) was used to sequence each sample to a depth of between 58-700-fold coverage (Table S1). The short reads (36-39bp) generated from the Solexa Genome Analyzer were mapped to the reference *Buchnera*-Ap sequence, Tokyo1998 (GenBank: NC_002528), and deviations from the sequenced strains (SNPs and small indels) were recorded (see below). For positions where we suspected, due to sequence coverage reduction, that larger insertions or deletions occurred or multiple substitutions were present within a restricted region, Sanger sequencing was used to resolve the sequence. Through a comparison with >30 kb of PCR-amplified regions sequenced for one or more strain with conventional Sanger methodology, we demonstrated that the genomic sequences generated through our re-sequencing approach contain no detectable errors, as required for our goal of identifying all intraspecific changes among strains. To further show that our resequencing approach generates reproducible results, we sequenced two independent samples of one strain, 8-10-1, that were obtained at different sampling dates in late 2007. Although these samples were processed and sequenced independently, we achieved a perfectly identical sequence of the full genome. Aphid strains are available from the corresponding author upon request.

Sequence reads were mapped to the *Buchnera aphidicola* Tokyo1998 genome with blastn with an e-value of 0.0001 and the –F F flag. These parameters allowed mapping of 39bp reads on the *Buchnera* genome with up to 5 mismatches per read. Only reads mapped by at least 33bp to the genome were taken into account. The alignment with the best bit-score was accepted as the correct mapping for each read, and in case a read had more than one best-scoring position on the genome it was multiply mapped and these positions were marked as repeats. For each sample, over 99.99% of the *Buchnera* reference genome was covered by uniquely mapped reads. Per-position coverage was calculated for each position with the mapped reads.

SNPs: Mismatches between individual reads and the Tokyo1998 genome were determined with the blast alignments. SNPs in the sequenced strain were called where a position was covered by at least 5 reads, with at least 80% of the covering reads showing the same mismatch. SNPs were categorized as replacements within coding regions, silent within coding regions, or occurring within intergenic spacers based on the annotation of the *Buchnera*-Ap Tokyo1998 genome.

Indels: Small indels were identified from the alignment between reads and the genome. Cases where an indel occurred within homopolymer were manually examined, as the blast alignment can place such indels at different positions along the homopolymer. Single base indels in homopolymers were the most frequent category of indels identified and could be unambiguously determined from the Solexa sequencing. Although indels in homopolymers are subject to high error rates in Sanger sequencing and particularly in 454 sequencing, they are not subject to elevated error or bias in Solexa sequencing (S2). Therefore indels of 1-3 bases could be reliably resolved by the set of individual Solexa reads spanning the homopolymer or other repetitive region. Since there is no initial amplification or cloning step, each Solexa read reflects an independent sample template. This precludes systematic error introduced during template amplification.

Large insertion/deletion events could be detected from alignments of Solexa reads on the reference genome; these gavea a signal of divergence from the reference but did not resolve the sequence. Deletions in a sequenced genome were determined as regions in the Tokyo1998 genome that were not covered by any single read. The short length of Solexa reads does not allow the direct sequence identification of large insertions in the sequenced genome relative to the reference genome. However, the positions of insertions could be detected, as reads aligned at these positions have a short unaligned tail that represents the beginning or end of the inserted sequence. Suspected insertion positions were therefore determined as positions where the alignment of reads to the Tokyo1998 genome was prematurely terminated in 80% or more of the covering reads. The exact sequence across each large insertion or deletion was determined by Sanger sequencing following PCR with primers based on regions flanking the detected insertion/deletion positions.

Sanger sequencing was used for selected regions and strains, to resolve ambiguous regions and to verify the accuracy of the Solexa-based resequencing. Over 30 kb of the 641 kb genome was verified with Sanger sequencing, for one to five of the sequenced strains. Changes in lab strains (5A-5AR) were verified as were all cases in which the Solexa sequences indicated a gene inactivation or larger indel. Primers were designed for flanking regions, PCR performed with standard conditions, and Sanger sequencing performed with an Applied BioSystems 3730 at the University of Arizona Genomic Analysis and Technology Center.

Genome sequences were deposited in GenBank with accession numbers CP00158 (for Tuc7 and 9-2-1) and CP00161 (for 5A, 5AR, A2A, 8-10-1, 7A). Variations within the two clusters were annotated as polymorphisms within these files.

Calculation of profile of new base changes and frequencies of silent (synonymous) and replacement (nonsynonymous) changes per site: To determine the relative frequencies of different mutation types, the tree was rooted with Tokyo1998, and directional changes were estimated for the two branches leading to Clusters 1 and 2 for the 6 types of base substitution (AT>CG, AT>GC, AT>TA, CG>AT, CG>TA, CG>GC). We then estimated the proportions of new mutations in coding regions that would result in silent changes and in replacement changes, by applying the observed mutational spectrum to each codon and weighting by the codon frequency in the *Buchnera*-Ap genome. First, each of the 9 base changes that could affect a codon were scored as silent or replacement and weighted for the frequency of that mutational category, to yield a codon-specific score for the ratio of replacement to silent changes among new mutations. This score was then weighted by the frequency of each codon in the *Buchnera*-Ap genome, and the codon-specific contributions were summed to give the genome-wide ratio of replacement to silent changes among new mutations, estimated at 4.51:1. This is the ratio of changes expected in the absence of selection. Of the 1996 base substitutions in coding regions, 718 were replacement changes and 1278 were silent changes, giving a ratio of replacement changes per replacement site to silent changes per silent site (equivalent to dN/dS) of 0.125.

Calibration of rates: *A. pisum* was first recorded in North American in Kansas in 1877 and spread across the US within a few years, based on multiple records from numerous regions in the US (S3, S4). We chose an initial date of 135 years before the time of the study (2007) to allow for the likelihood that the actual introduction was a few years earlier and to ensure that the date gave an upper bound on the coalescence time of our strains. The initial rapid expansion in population size and geographic range is likely to result in the persistence of lineages diverging from the initial colonizers, even if later introductions took place. The greater incidence of transoceanic trade during the 20th century would likely result in more introductions and in reverse introductions (US to Europe) (S3). Cluster 1 and Cluster 2 from Fig. 1 were inferred to represent two separate introductions to North America on the basis of their relationships to worldwide collections. Evidence for this conclusion was from wider sampling of *Buchnera*- Ap with collections taken between 1988 and 2007 from localities in North America (AK, CA, NY, OR, SC, UT, WI, NY) and several samples from Europe and East Asia. A region of intergenic spacer corresponding to positions 30175-31218 in the reference sequence (NC_002528) was sequenced for 37 collections. Variable sites were used to construct a haplotype tree, shown in Fig. S2. Cluster 2 is exclusively North American. Cluster 1 is largely North American but cannot be separated from some European samples, on the basis of this limited sequence region.

Calculation of rates of nucleotide substitution: The two North American clusters yielded similar average divergences, as discussed in the results. In order to calculate the variance of the rate estimate, a Poisson distribution of substitutions was assumed, lumping all of the changes within each of the two clusters. The pooled data gave a total of 38 substitutions/641 kb/540 years. The 95% confidence interval was calculated on the basis of the Poisson approximation of a binomial distribution of changes on sites in the genome. Because some changes have been subject to purifying selection, this rate is expected to be lower than the spontaneous rate of new mutations. We thus calculated the rate using only changes in spacers and synonymous changes. The proportion of new mutations that are synonymous was calculated as described above. These were then pooled with spacer sites to give 209072 sites. The neutral substitution rate was calculated as 22 substitutions/209 kb/540 years.

Ancestral ages for divergences among the three main lineages were calculated by applying calibrated rates to the observed divergence. The deeper divergence was the average for Tokyo1998-Cluster 1 and Tokyo1998-Cluster 2, the other divergence was for Cluster 1 - Cluster2. Dates were calculated with all sites in the genome for both calibration and dating and with only spacer and synonymous changes for both calibration and dating.

Table S1. Genes inactivated in some sequenced genomes of *Buchnera*-Ap and the nature of the inactivating mutation.

gene	genomic position of change (Tokyo1998)	product	function	samples-gene intact*	samples-gene not intact	Basis for inactivation
rnpA	14613	ribonuclease P protein component	RNA processing	Tokyo1998	C1, C2	indel in homopolymer
yigL	29970	hypothetical sugar phosphatase	hydrolase activity	C1, C2	Tokyo1998	indel
argC	52972	N-acetyl-gamma-glutamyl-phosphate reductase	amino acid biosynthesis	All except 8-10-1	8-10-1 only	indel
fliK	82532	flagellar hook-length control protein FliK	protein secretion apparatus	C1, C2	Tokyo1998	substitution causing stop codon
ung	199877	uracil-DNA glycosylase	DNA repair	Tokyo1998, C1	C2	indel in homopolymer
hpt	212855	hypoxanthine phosphoribosyl-transferase	nucleotide metabolism	C1, C2	Tokyo1998	substitution causing stop codon
murF	242374	UDP-N-acetylmuramoylalanyl-D-glutamyl-2,6-diaminopimelate-D-alanyl-D-alanyl ligase	cell wall biogenesis	Tokyo1998	C1, C2	indel in homopolymer
fabZ	262608	6-diaminopimelate-D-alanyl-D-alanyl ligase	fatty acid biosynthesis	C1, C2	Tokyo1998	large deletion removing entire gene
yaeT	262608	subunit of Outer Membrane Protein Assembly Complex	Outer membrane component	C1, C2	Tokyo1998	large deletion removing 36% of coding region
trpB	302801	tryptophan synthase, beta subunit	amino acid biosynthesis	Tokyo1998	C1, C2	indel
glyA	318768	serine hydroxymethyl-transferase	amino acid biosynthesis	Tokyo1998	C1, C2	indel in homopolymer

flgB	370630	flagellar basal-body rod protein FlgB	protein secretion apparatus	Tokyo1998	C1, C2	indel
recC	497229	exodeoxyribonuclease V 125 kDa polypeptide	DNA repair	C1, C2	Tokyo1998	substitution causing stop codon
ribD2	508064	riboflavin reductase	flavin biosynthesis	Tokyo1998, C1	C2	indel in homopolymer
sbcB	590197	exonuclease I, 3' --> 5' specific, deoxyribophosphodiesterase	DNA repair	Tokyo1998	C1, C2	indel in homopolymer
ynfM	621255	YnfM MFS transporter	membrane transporter	Tokyo1998	C1, C2	indel in homopolymer

*Tokyo1998 is reference sequence, C1 includes Tuc7 and 9-2-1, C2 includes 5A, 5AR, 7A, 8-10-1, A2A

Fig. S1. Chromosomal positions of nucleotide substitutions and indels in eight fully sequenced strains of *Buchnera*-Ap. Base substitutions and indels are approximately randomly distributed across the genome aside from a strong concentration of indels in intergenic spacers.

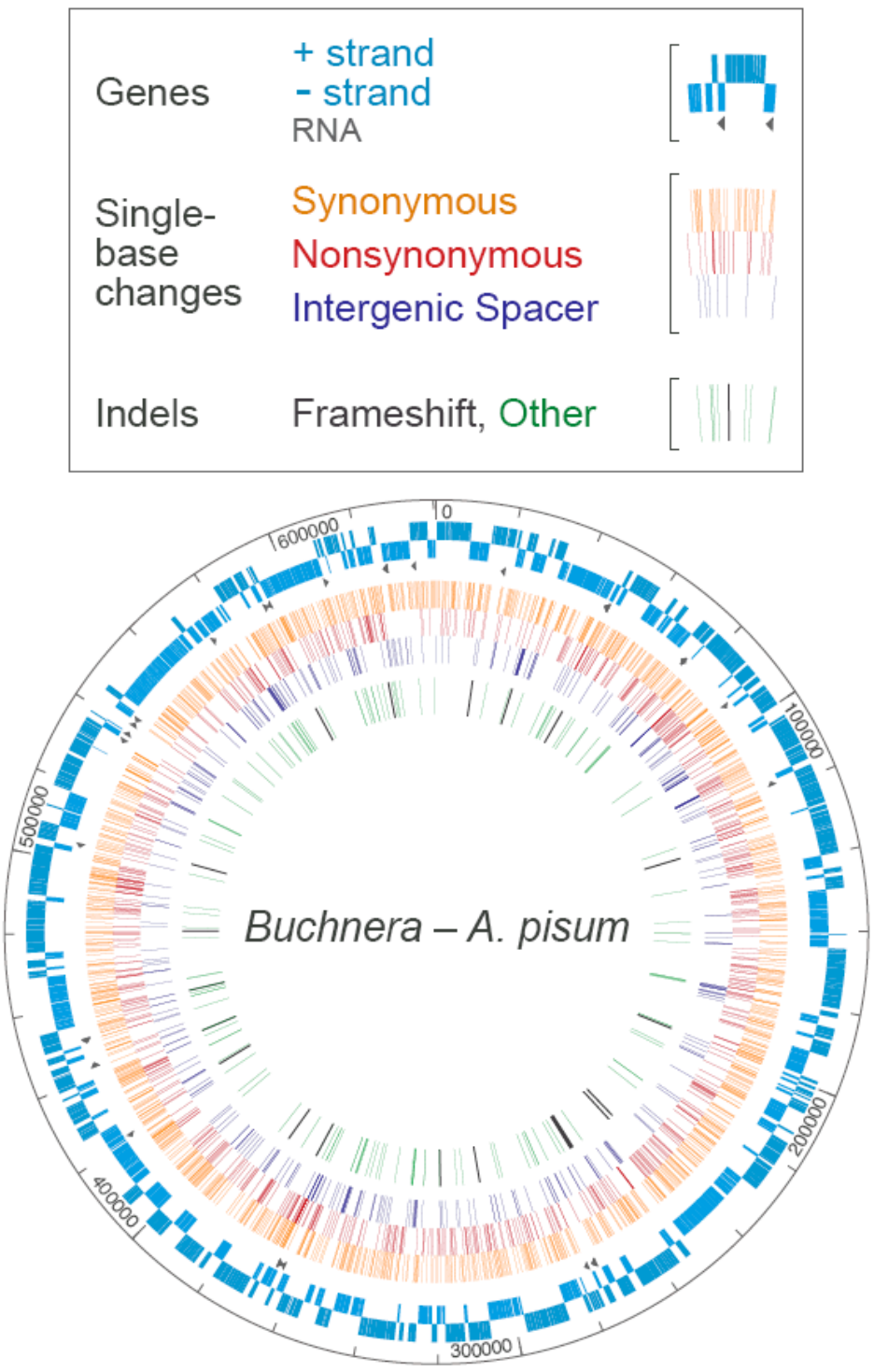
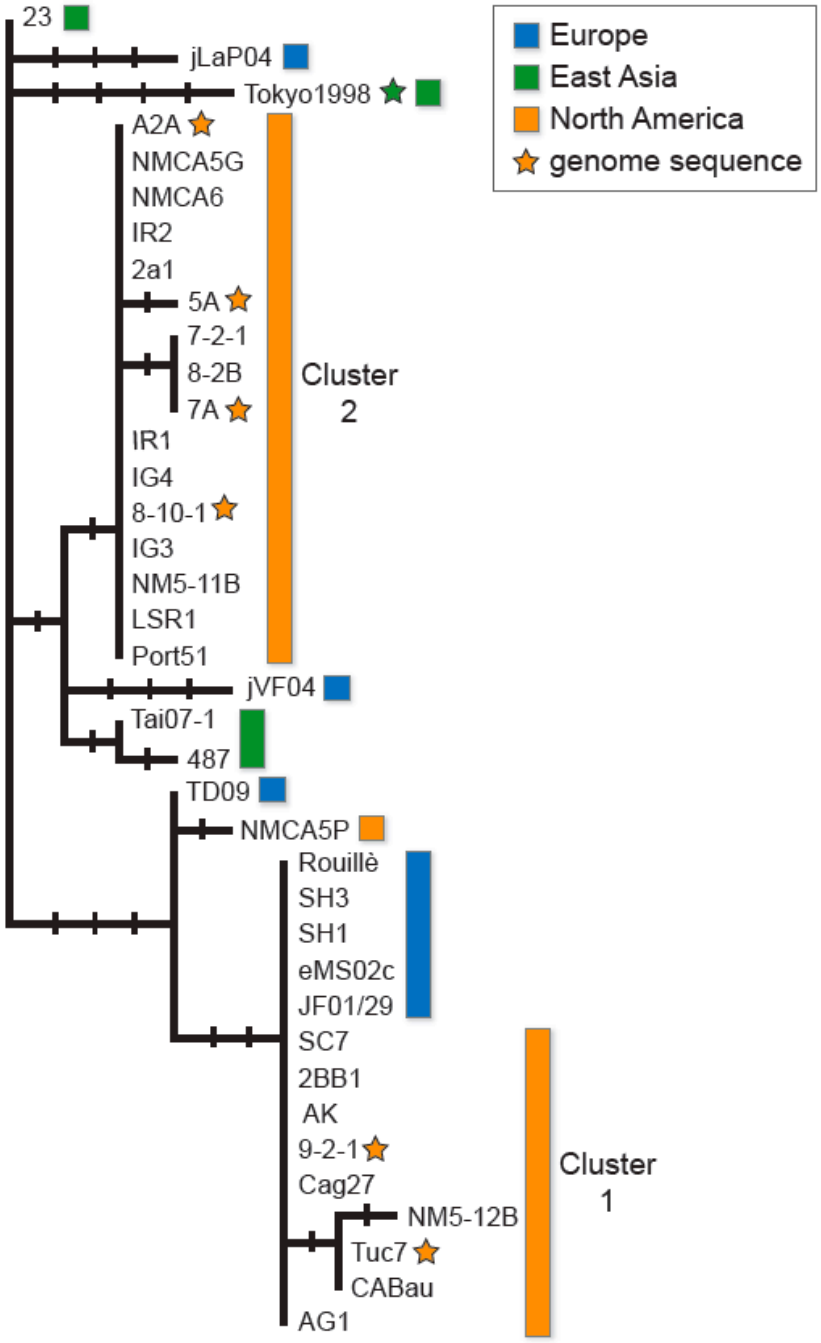


Fig S2. Haplotype tree for *Buchnera*-Ap based on 1045 base pairs (positions 30175-31218 in the Tokyo1998 reference sequence, NC_002528). Cluster 2 is exclusively North American and represented by samples from CA, OR, UT, WI, NY. Cluster 1 is largely North American and represented by samples from AK, CA, NY, SC, WI.



Supporting References and Notes

S1. I. Tamas *et al*., Fifty million years of genomic stasis in endosymbiotic bacteria. *Science* **296**, 2376-79. (2002).

S2. D. R. Bentley *et al.*, *Nature* **456**, 53-59. (2008)

S3. R. G. Foottit, S. E. Halbert, G. L. Miller, E. Maw, L. M Russell, *Entomol. Soc. Wash.* **108**, 583-610. (2006).

S4. A. M. Harper, J. P. Miska, G. R. Manglitz, B. J. Irwin, E. J. Armbrust, A bibliography of the pea aphid *Acyrthosiphon pisum* (Harris) (Homoptera: Aphididae). (Special publication 50, University of Illinois, Agricultural Experimental Station, Urbana-Champaign. pp. 1-89 (1978).