Extremophiles (2008) 12:365–374
DOI 10.1007/s00792-008-0141-2

ORIGINAL PAPER

Reductive genome evolution in chemoautotrophic intracellular symbionts of deep-sea *Calyptogena* clams

Hirokazu Kuwahara · Yoshihiro Takaki · Takao Yoshida · Shigeru Shimamura · Kiyotaka Takishita · James D. Reimer · Chiaki Kato · Tadashi Maruyama

Received: 23 October 2007 / Accepted: 7 January 2008 / Published online: 28 February 2008

Abstract To understand reductive genome evolution (RGE), we comparatively analyzed the recently reported small genomes of two chemoautotrophic, intracellular symbionts of deep-sea clams, *Calyptogena okutanii* and *C. magnifica*. Both genomes lack most genes for DNA recombination and repair such as *recA* and *mutY*. Their genome architectures were highly conserved except one inversion. Many deletions from small (<100 bp) to large (1–11 kbp) sizes were detected and the deletion numbers decreased exponentially with size. Densities of deletions and short-repeats, as well as A+T content were higher in non-coding regions than in coding regions. Because *Calyptogena* symbiont genomes lack *recA*, we propose that deletions and the single inversion occurred by RecA-independent recombination (RIR) at short-repeats with simultaneous consumption of repeats, and that short-repeats were regenerated by accelerated mutations with enhanced A+T bias due to the absence of *mutY*. We further propose that extant *Calyptogena* symbiont genomes are in an actively reducing stage of RGE consisting of small and large deletions, and the deletions are caused by short-repeat dependent RIR along with regeneration of short-repeats. In future, the RGE rate will slowdown when the gene repertoires approach the minimum gene set necessary for intracellular symbiotic life.

Keywords Reductive genome evolution (RGE) · *recA* · Comparative genomics · Deletion · Repeat · A+T bias · *Calyptogena* · Symbiosis

Introduction

Intracellular symbionts provide insights into reductive genome evolution (RGE). They appear to have evolved by reducing their genomes from their ancestors. It has been proposed that the evolution of bacterial genome size depends on two opposing forces, DNA acquisition and its loss, and that deletion bias is the major force that shapes their genomes (Mira et al. 2001). Because intracellular symbionts are sheltered by host cells, they are thought to have less opportunity to incorporate foreign DNA (Mira et al. 2001). RGE has been extensively studied in obligate, intracellular, maternally transmitted symbionts of insects (Delmotte et al. 2006; Klasson and Andersson 2004; Moran 2003; Tamas et al. 2002; Moran and Mira 2001). Genomes of two strains of *Buchnera aphidicola*, aphid symbionts, have been shown to have similar sizes and conserved gene orders, which are likely attributable to loss of *recA* as well as phages and repeated sequences (Tamas et al. 2002). It has been proposed that these

Communicated by K. Horikoshi.

H. Kuwahara and Y. Takaki equally contributed to this work.

Electronic supplementary material The online version of this article (doi:10.1007/s00792-008-0141-2) contains supplementary material, which is available to authorized users.

H. Kuwahara · Y. Takaki · T. Yoshida · S. Shimamura · K. Takishita · J. D. Reimer · C. Kato · T. Maruyama (✉)
Extremobiosphere Research Center,
Japan Agency for Marine-Earth Science and Technology
(JAMSTEC), 2-15 Natsushima-cho,
Yokosuka 237-0061, Japan
e-mail: tadashim@jamstec.go.jp

J. D. Reimer
Department of Marine Science, Biology and Chemistry,
University of the Ryukyus, Senbaru 1,
Nishihara, Okinawa 903-0213, Japan

symbionts have lost *recA* as a consequence of relaxed selection in the early stage of genome reduction (Dale et al. 2003). In modern *Buchnera* strains the gene repertoires of their genomes are thought to be approaching the minimum gene set which are necessary for their intracellular symbiotic lives, and thus RGE is in a relatively stable stage (Klasson and Andersson 2004; Moran 2003). Large size deletions spanning multiple genes have been proposed to occur in the early phases of RGE after the establishment of intracellular symbiosis (Delmotte et al. 2006; Klasson and Andersson 2004; Mira et al. 2001; Moran 2003; Moran and Mira 2001). However, it is not clear whether the loss of *recA* and the occurrence of large DNA deletions also happen during the early-phase processes of RGE in other intracellular symbionts. Also the details of the deletion mechanisms of RGE, e.g., roles of *recA* and of repeated sequences in the large deletion events, remain to be studied. To understand the general process of RGE and mechanisms of the actively reducing stage of RGE, it is necessary to examine the genomes of intracellular symbionts other than insect symbionts in the stable stage of RGE.

Recently, the genomes of two chemoautotrophic intracellular symbionts of deep-sea clams, symbionts of *Calyptogena okutanii* (Candidatus *Vesicomyosocius okutanii*: Vok, found in host clams at seeps) and of *C. magnifica* (Candidatus *Ruthia magnifica*: Rma, found in host clams at hydrothermal vents), have been reported (Kuwahara et al. 2007; Newton et al. 2007). The symbionts are vertically transmitted via eggs (Endow and Ohta 1990) and are thought to have co-evolved with the *Calyptogena* clams (Peek et al. 1998).

The clam symbiont genomes have genes for chemoautotrophic metabolism, but lack most genes for DNA recombination and repair, such as *recA* and *mutY* (Kuwahara et al. 2007; Newton et al. 2007). The Vok and Rma genomes are small (1.02 and 1.16 Mb, respectively) and have high A+T content. Their genomes are smaller than those of their free-living relatives and have been suggested to have been reduced from their common ancestor's genome. Considering their close phylogenetic relationship (Electronic supplementary material, Fig. S1) the difference between their genome sizes is unexpectedly large. This difference suggested that the genomes of *Calyptogena* symbionts are in the actively reducing stages of RGE. Comparison of these two genomes may allow us to gain insight into the processes of RGE. Here, we show that in modern *Calyptogena* symbiont genomes both small and large deletions are ongoing and are mediated by short-repeat dependent but RecA-independent recombination. We also present a broader perspective on the nature of RGE in intracellular symbionts.

Materials and methods

Genome sequence data

The complete genome sequences with annotations of *Calyptogena okutanii* symbiont (Candidatus *Vesiomyosocius okutanii*: Vok), *C. magnifica* symbiont (Candidatus *Ruthia magnifica*: Rma), *Buchnera aphidicola* APS, *B. aphidicola* Sg, *Blochmannia floridanus*, *B. pennsylvanicus*, *Escherichia coli* K12, *Salmonella typhimurium* (accession numbers: NC_009465, NC_008610, NC_002528, NC_004061, NC_005061, NC_007292, NC_000913 and NC_003197, respectively) were downloaded from the NCBI Entrez website.

Alignment of whole genomes

The whole genomes of symbionts were aligned using a global alignment program, LAGAN, with default parameters (Brudno et al. 2003). Conserved regions were defined as sequences of over 100 bp with over 70% homology, and used as anchor points for further alignments. Gaps larger than 10 bp found in the alignment were designated as deletions while gaps less than 10 bp were not.

Searching method for repeated sequences

Repeated sequences larger than 8 bp were searched by using Repseek program (Achaz et al. 2007). We searched for repeated sequences within a window of 1,000 bp. Long repeats (>200 bp) were also searched with the Repseek program over the whole genome sequence.

Identification of orthologous genes and pseudogenes

Orthologous groups between each genome pair were classified using InParanoid program (O'Brien et al. 2005). First, all possible pairwise similarity scores that scored higher than a cutoff value (bitscore ≥50, overlap ≥70%) were detected from all-against-all BLAST (Altschul et al. 1997) comparisons and then the reciprocal genome-specific best hits were marked as orthologs. MultiParanoid program (Alexeyenko et al. 2006) was used in order to classify orthologous groups between protein coding genes in multiple genomes. This program applies a clustering algorithm to merge multiple-pairwise orthologous groups from InParanoid into multi-species orthologous groups. Through the Inparanoid algorithm, 857 ortholog pairs were identified between *Calyptogena* symbiont protein-coding genes and 317 ortholog groups were common in all protein-coding genes of *Calyptogena* symbionts, *Buchnera aphidicola* strains, *Blochmannia floridanus*, *B. pennsylvanicus*, *Escherichia coli* K12 and *Salmonella typhimurium*.

In order to identify pseudogenes, an intergenic region in each *Calyptogena* symbiont genome was subjected to BLASTX analysis against the protein database of the other. Pseudogenes were detected as segments with similarity to functional ORFs in other genomes, but with multiple indels and missense mutations resulting in frameshift or in-frame stop codon disruptions.

Estimate of synonymous and nonsynonymous substitution rates

Protein sequences of each pair of orthologs were aligned using ClustalW (Thompson et al. 1994) and nucleotide sequences were aligned according to the corresponding protein sequence alignment using CodonAlign 2.0 (Hall 2004). Pairwise estimates of the synonymous (*dS*) and nonsynonymous (*dN*) substitution rates were obtained by the modified Pamilo–Bianchi–Li method (Li 1993; Pamilo and Bianchi 1993; Tzeng et al. 2004), as implemented in the KaKs_Calculator program (Zhang et al. 2006).

Results and discussion

Deletions in *Calyptogena* symbiont genomes

The alignment of *Calyptogena* symbiont genomes showed that the two-genome sequences are highly homologous (Table 1) and that their gene orders (syntenies) are highly conserved except for one inversion (Fig. 1). Their genome structural stability is probably due to the lack of DNA recombination genes, including *recA* (Table 2) (Kuwahara et al. 2007; Newton et al. 2007). Recombination is the major source of genome reorganization. Two types of recombinations are known: RecA-dependent homologous recombination requiring large-repeats (>200 bp) and RecA-independent recombination (RIR) requiring short-repeats (>8 bp) (Lovett 2004; Rocha 2003). The lack of *recA* and the absence of exact repeats larger than 200 bp in the genomes of both lineages of *Calyptogena* symbionts suggest that the inversion occurred in one of the genomes by RIR after their divergence (Bzymek and Lovett 2001; Lovett 2004).

In contrast to the conserved synteny, many gaps of varying sizes (up to 11 kbp) were scattered within each of the two genomes (Fig. 1). The number of such gaps was twice as high in Vok compared to Rma (Fig. 1; Table 1). In *Calyptogena* symbiont genomes, we found neither mobile elements (e.g., transposons) nor sequences that could be regarded as exogenous DNA (obtained by horizontal transfer). In an intracellular lifestyle, the opportunity for gene uptake is assumed to be reduced (Mira et al. 2001). Therefore, these gaps are mostly attributable to deletions (Table 1; Fig. 1).

The number of deletions was found to decrease exponentially with deletion size (Fig. 2). The majority of deletions were less than 100 bp in length (77% of the total deletion number), but the occupancy (21.5% of the total length) was lower. The summed length of moderate-sized

Table 1 Comparison of two *Calyptogena* symbiont genomes

	Calyptogena okutanii symbiont (Vok)	*Calyptogena magnifica* symbiont (Rma)
Genome size (bp)	1,022,154	1,160,782
G+C content (%)	31.6	34.0
Number of protein coding genes	937	976
Non-coding regions (%)	14.2	20.2
rRNA operons (number)	1	1
tRNA genes (number)	35	36
Average gene length (bp)	928	943
Average intergenic length (bp)	149	232
Number of orthologs (% of the length in the genome size)	857 (79.8)	857 (70.3)
Averaged homology of orthologs between the two genomes (%)	82.1	82.1
Average homology of intergenic regions between two genomes (>100 bp)	74.8	74.8
Number of deletions (>10 bp consecutive gaps)	1,387	730
Size range of the deletions (bp)	11–10,964	11–1,755
Total length of deletions (bp) (% of the length in total deletion length)	195,460 (19.1)	56,378 (4.8)

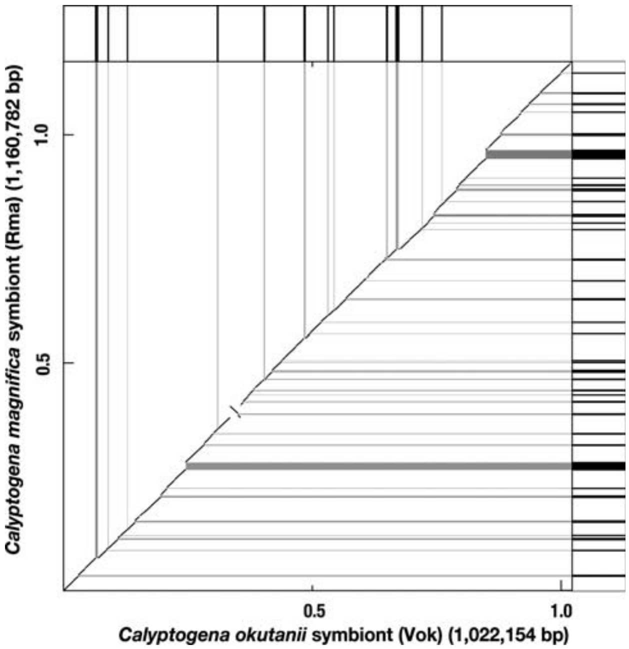


Fig. 1 Genome alignment of *Calyptogena okutanii* symbiont (Candidatus *Vesicomyosocius okutanii*: Vok) and *C. magnifica* symbiont (Candidatus *Ruthia magnifica*: Rma). Gaps (≥500 bp with homology less than 70%) are shown in *black bars* on the right and upper axes. Note inversion in the middle of the genomes [Vok 335,864–355,593 (19,730 bp); Rma 380,305–404,033 (23,729 bp)]

(from 100 to 1,000 bp) deletions made up 50.3% of total deletions and the large deletions (≥1,000 bp) 28.2% of total deletions (Fig. 2). These data indicate that not only small and moderate-sized deletions but also deletions larger than one gene in size (928–943 bp; Table 1) have significantly contributed to RGE.

In both *Calyptogena* symbiont genomes, the majority of deletions were found within non-coding regions (70.1% in total number and 56.9% in total length) as opposed to deletions in coding regions (29.9% in total number, 43.1% in total length) (Fig. 2). Considering the relatively low percentage of non-coding regions in both genomes (14.2 for Vok and 20.2 % for Rma in size; Table 1), these results suggest that deletions have occurred more frequently in non-coding regions than in coding regions.

Extant *B. aphidicola* strains, APS and Sg, are thought to be in the late-stable stage of RGE (Klasson and Andersson 2004; Tamas et al. 2002). The deletions in their aligned genome sequences were found to be mostly less than 100 bp (95.1 in number, 67.4 % in size; maximum size of single deletion is 564 bp). In addition, extant *Blochmannia floridanus* and *B. pennsylvanicus*, which are symbionts of carpenter ants, have different genome sizes of 706 and 792 kb, respectively (Gil et al. 2003; Degnan et al. 2005), but show complete conservation of genome architecture (Degnan et al. 2005). Based on aligned genome sequences of *Blochmannia* species, deletions shorter than 100 bp and those between 101 and 500 bp were found to occupy 91.0 and 8.7% of the total deletion numbers, respectively. The summed deletion lengths of these deletions made up 60.4 and 34.8% of the total deletion length, respectively. Only one deletion was found to be larger than 1,000 bp (1,169 bp).

Table 2 Genes for DNA repair and recombination in genomes of *Calyptogena okutanii* symbiont (*Vesicomyosocius okutanii*: Vok), *C. magnifica* symbiont (*Ruthia magnifica*: Rma), *Buchnera aphidicola* APS (Buc), *B. aphidicola* Sg (Bas), *Blochmannia floridanus* (Bfl), *B. pennsylvanicus* (Bpn), *Escherichia coli* (Eco), and *Salmonella typhimurium* (Stm)

Function	Gene	Vok	Rma	Buc	Bas	Bfl	Bpn	Eco	Stm
Direct damage reversal	*phrB*	–	–	+	–	–	–	+	+
	ada	–	–	–	-	–	–	+	+
	ogt	–	–	–	–	–	–	+	+
Base excision repair	*ung*	+	+	+	–	+	+	+	+
	tag	–	–	–	–	–	–	+	+
	alkA	–	–	–	–	–	–	+	+
	mutM	+	+	–	–	–	–	+	+
	mutY	–	–	+	+	+	+	+	+
	nth	+	+	+	+	+	+	+	+
	nfo	–	–	+	+	–	–	+	+
Mismatch repair	*mutS*	–	–	+	+	–	–	+	+
	mutL	–	–	+	+	–	–	+	+
	mutH	–	–	–	–	–	–	+	+
	recJ	+	+	–	–	–	–	+	+
	uvrD	+	+	–	–	–	–	+	+
Oxidative damage repair	*mutT*	–	+	+	+	–	–	+	+
Recombinase pathways	*recA*	–	–	–	–	–	–	+	+
	recB	–	–	+	+	+	+	+	+
	recC	–	–	+	+	+	+	+	+
	recD	–	–	+	+	+	+	+	+
	recF	–	–	–	–	–	–	+	+
	recN	–	–	–	–	–	–	+	+
Nucleotide excision repair	*uvrA*	+	+	–	–	–	–	+	+
	uvrB	–	+	–	–	–	–	+	+
	uvrC	–	+	–	–	–	–	+	+
	mfd	+	+	+	–	–	–	+	+

+ Presence of the gene; – absence of the gene

These data suggest that the *Calyptogena* symbionts are in an earlier, actively reducing stage of RGE with larger deletions than *B. aphidicola* and *Blochmannia* species.

Possible mechanism of reductive genome evolution (RGE) in *Calyptogena* symbionts

Deletions were found in clusters in both non-coding and coding regions (Fig. 3a, b). Small deletions likely took place randomly but more frequently in regions with lower

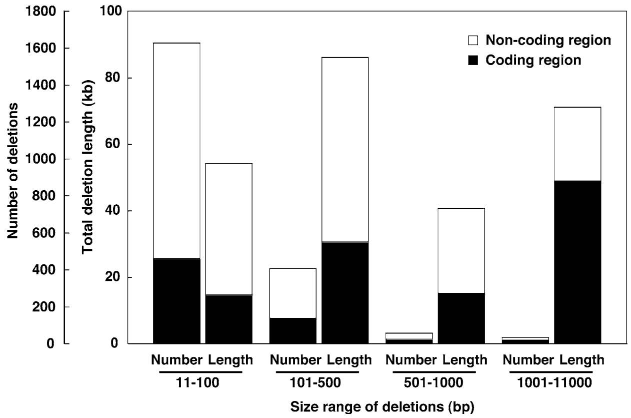


Fig. 2 Size distribution of the deletions in both *Calyptogena* symbiont genomes. Number, total number of deletions for each size range of deletions. Length, summed length of deletions of each size range. *Open-area* of the *bar*, deletions in non-coding regions; *filled-area* of the *bar*, deletions in coding regions

functional constraints, such as non-coding regions. As a result, small deletions accumulated and formed apparent larger deletions or deletion clusters, and contributed to RGE. Some pseudogenes were found in non-coding regions (Fig. 3b). This suggests that some non-essential genes accumulated point or frameshift mutations, and changed to non-coding regions via pseudogenes, which were then lost by multiple small deletions (Fig. 3b). In addition, deletions were frequently found in the flanking regions of tRNA genes (Fig. 3c). Transfer RNA genes have been suggested to be the loci for gene rearrangement in *Prochlorococcus* (Rocap et al. 2003). Transfer RNA genes may have played a role in the RGE of *Calyptogena* symbiont genomes as repeats. In contrast to small-sized deletions, large-sized deletions often contained several coding genes (Figs. 2, 3d). By such large-size deletions, coding regions containing non-essential genes were likely to be lost by single deletion events. In extant *Calyptogena* symbionts, RGE seems to have proceeded mostly by small-sized deletions and less frequently by larger deletions of blocks of non-essential genes.

In RIR, at least two different recombination mechanisms are known; DNA replication slippage (DRS) and single-strand annealing (SSA), both of which take place between short-repeats at short distances apart (Lovett 2004; Rocha 2003). SSA does not seem to have contributed to RGE of *Calyptogena* symbionts, because genes encoding SSA-proteins such as RecT and RAD52 (Iyer et al. 2002) were not found in their genomes. In DRS the deletion rate is expected to depend exponentially on proximity between repeats (>8 bp), and deletion lengths can be up to approximately 10 kbp (Lovett 2004). In *Calyptogena* symbiont genomes, the number of deletions was found to decrease exponentially with size and the maximum deletion size was 10,964 bp (Fig. 2). Furthermore, the entire genome and especially non-coding regions were overlaid with a high density of short repeats (>8 bp) (Figs. 4, 5). These data support the view that the deletions in *Calyptogena* symbionts may have occurred by DRS.

RecA-dependent homologous recombination and DRS are known to consume large (≥200 bp) and short repeats (≥8 bp), respectively (Frank et al. 2002; Lovett 2004). In *Calyptogena* symbiont genomes, we found any exact repeat larger than 200 bp but two pairs of paralogous genes in each genome (hypothetical protein; Rmag_0424/Rmag_1074, COSY_0393/COSY_0973; translation elongation factor EF-Tu; Rmag_0163/Rmag_0818, COSY_0167/COSY_0744). In various bacteria, the average density of identical repeats >200 bp has been reported to be 1.7% per genome (Frank et al. 2002). If we postulate that the *Calyptogena* symbionts had *recA* in the initial stages of RGE, the absence of large repeats (>200 bp) in Vok and Rma may indicate that they have been consumed by RecA-dependent homologous recombination. After the consumption of short-repeats during DRS, regeneration of short-repeats is necessary for further RGE to occur. In *Calyptogena* symbionts, the lack of DNA repair genes, including *mutY* (Table 2) (Kuwahara et al. 2007; Newton et al. 2007) probably increased the mutation rate, especially in non-coding regions where functional constraints were lower than coding regions. In addition, the lack of *mutY* is known to increase A+T bias at mutations (Au et al. 1989; Nghiem et al. 1988). The higher mutation rate with mutational bias towards higher A+T content is expected to decrease the complexity of the sequence and increase the probability of producing short-repeats, especially in non-coding regions. Indeed, densities of repeats and A+T content are both higher in non-coding regions than in coding regions (Figs. 4, 6). These data indicate that the higher mutation rate with higher A+T bias is a driving force for deletions in RGE of the extant *Calyptogena* symbionts.

It may be noteworthy that the smaller genome of Vok lacks *uvrB* and *uvrC* that suppress RIR (Hanada et al. 2000), but Rma has *uvrB* and *uvrC*. Spontaneous deletion of the genes for DNA repair and recombination might affect the rate of RGE in the later stage. Loss of *uvrB* and *uvrC* may be one mechanism causing the difference in sizes between the two *Calyptogena* symbiont genomes. It seems that RGE of *Calyptogena* symbionts is ongoing and *uvrB* and *uvrC* may be lost in Rma in the future.

Relaxed functional constraint and mutation rate of *Calyptogena* symbiont genes

The rates of nonsynonymous and synonymous substitutions (*dN* and *dS*) within four “pairs”; two *Calyptogena*

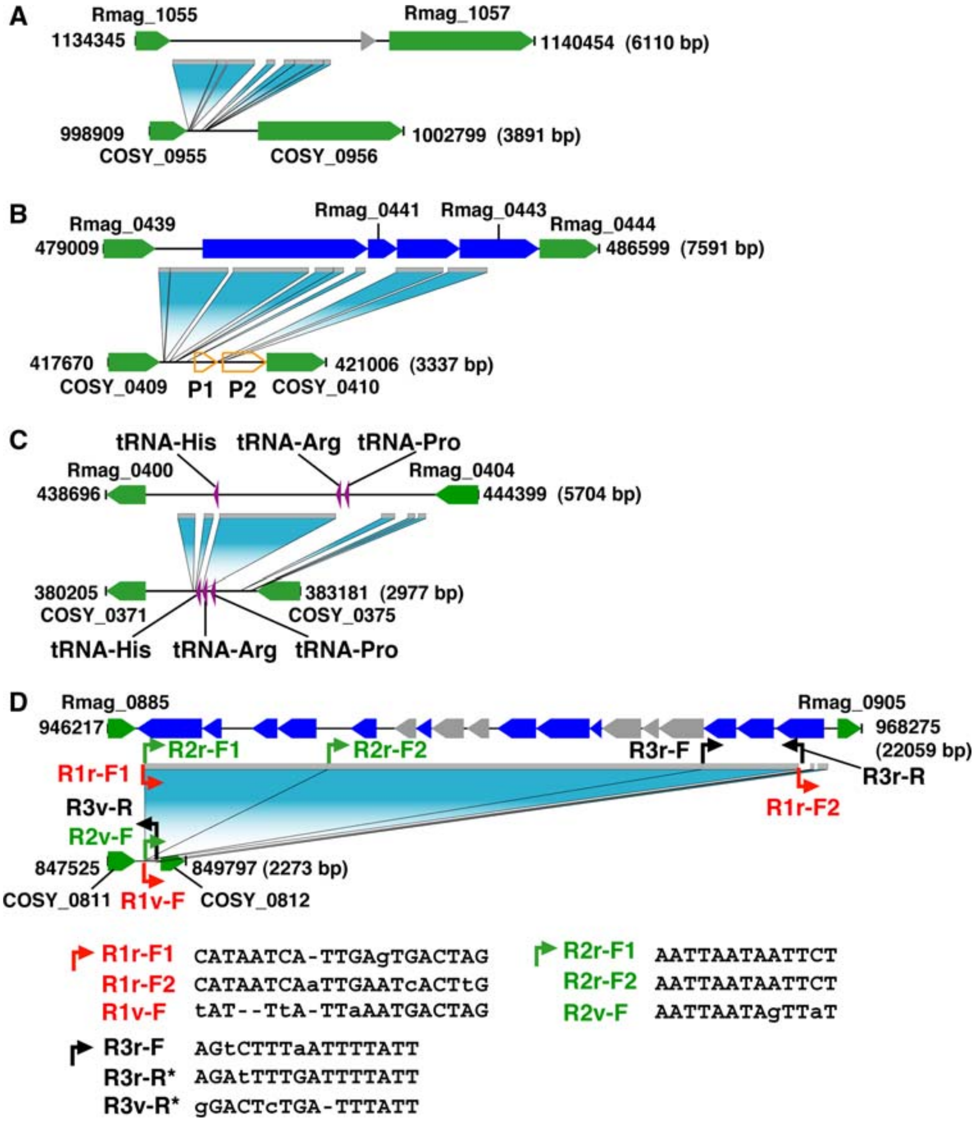


Fig. 3 Map of deleted genes and corresponding remaining sequences in parts of *Calyptogena* symbiont genomes. Upper lines with *pentagonal columns*, *C. magnifica* symbiont (*Ruthia magnifica*: Rma) genome. Lower lines with *pentagonal columns*, *C. okutanii* symbiont (*Vesicomyosocius okutanii*: Vok) genome. *Numbers* at the terminals of the sequences, nucleotide number from the replication origin. *Numbers* with Rmag or COSY, gene ID of Rma or Vok. *Pentagonal column*, protein coding gene. *Green pentagonal column*, conserved ortholog. *Yellow column*, protein coding gene deleted in the Vok genome. *Gray pentagonal column*, hypothetical gene. *Gray-line* under the upper line with the *blue shadowed area*, the sequences deleted from Vok genome. **a** Deletions in a non-coding region. Repeated sequences in this area of Rma genome are shown in Fig. 5. **b** Deletions in a coding region. *Orange pentagonal open columns*, pseudogenes in Vok genome. **c** Deletions in the flanking regions of tRNA genes. *Violet arrowhead*, tRNA genes. **d** A apparent large size deletion (19.8 kbp) containing coding regions. Relatively long repeat-like sequences very close to the terminals of deleted area are shown. The large deletion between R1r-F1 and R1r-F2 is composed of three flanking deletions, which are separated with short sequences homologous to the corresponding sequences of the Vok genome. The largest deletion is at position Rma 952663–963626 (length 10,964 bp) between repeats, R2r-F2 and R3r-F. This is the largest deletion detected in the Vok genome. The second largest deletion is at the position of Rma 947291–952653 (length 5,363 bp) between the repeats R2r-F1 and R2r-F2. The smallest deletion is at Rma 963637–966433 (length 2,797 bp) between repeats R3r-F and R1r-F2. Nucleotide sequences of the repeats are shown but the sequences of R3r-R* and R3v-R* are shown as their complementary sequences

symbionts, two *Buchnera aphidicola* strains, two *Blochmannia* species, and *Eschericha coli–Salmonella typhimurium*, were calculated using 317 orthologs. The values of *dN*/*dS* of the two *Calyptogena* symbionts (0.20 ± 0.09), *B. aphidicola* strains (0.19 ± 0.11) and *Blochmannia* species (0.25 ± 0.17) were similar and 4–5 times higher than *dN*/*dS* of *E. coli*–*S. typhimurium* (0.05 ± 0.08) (Table 3). These results indicate that functional constraints of genes in the symbionts have been relaxed, probably due to intracellular lifestyle. Furthermore, the lack of DNA

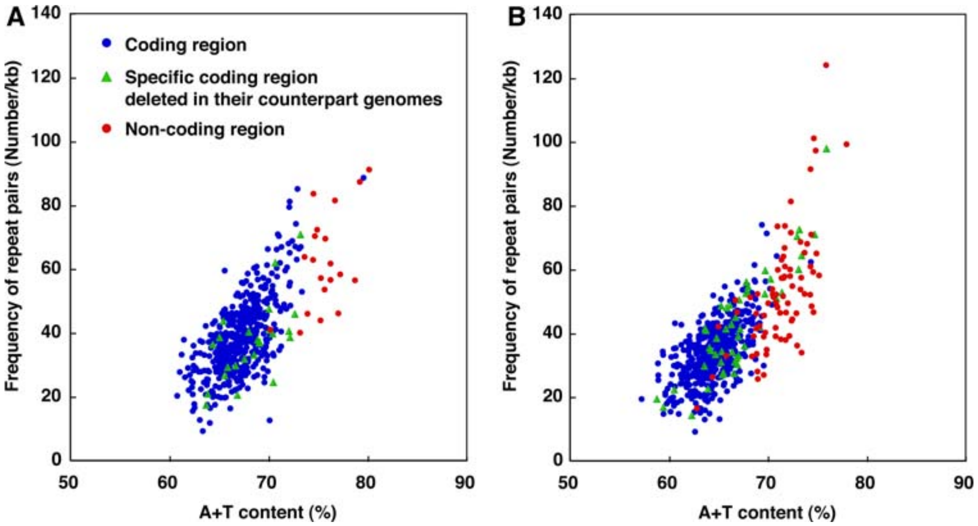


Fig. 4 Frequencies of repeat pairs in the *Calyptogena* symbiont genomes. Repeats larger than 8 bp were searched for in the genomes of *Calyptogena* symbionts; **a** in *C. okutanii* symbiont (*Vesicomyosocius okutanii*: Vok), **b** in *C. magnifica* symbiont (*Ruthia magnifica*: Rma). Repeats were counted in coding and non-coding regions larger than 1,000 bp. They were also counted in Vok or Rma-specific coding regions (>1,000 bp), which were deleted in their counterpart genomes. Frequency of the pairs of repeat per 1,000 bp sequence was plotted against the A+T content of the sequence (>1,000 bp). The frequency of repeats and the A+T content were found to be higher in non-coding regions than in coding region

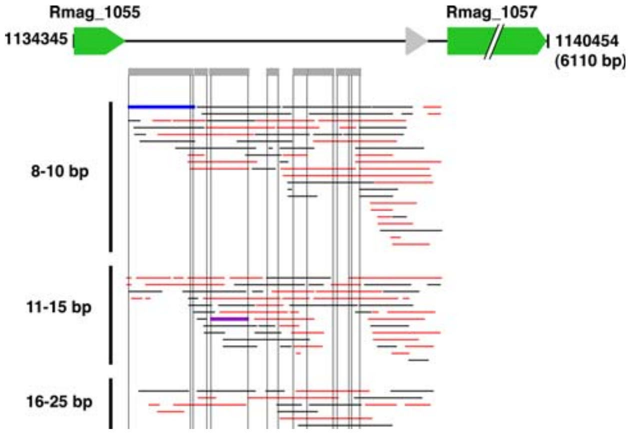


Fig. 5 Map of short-repeats in a non-coding region. Upper line, portion of *Calyptogena magnifica* symbiont (*Ruthia magnifica*: Rma) genome sequence (the same region as in Fig. 3a). *Gray line*, sequences deleted from *C. okutanii* symbiont (*Vesicomyosocius okutanii*: Vok) genome. *Black* or *red lines* indicate possible deletions caused by the RecA-independent recombination (e.g., DNA replication slippage) between pairs of short-repeats. Direct and inverted repeats are shown as *black* and *red lines*, respectively. A pair of repeats locate at the terminals of each line. The *blue thick line* (direct repeat) and *purple thick line* (inverted repeat) represent possible deletions, which are very similar in length and positions to corresponding deleted regions from Vok. Repeats are shown in three size classes: 8–10, 11–15 and 16–25 bp

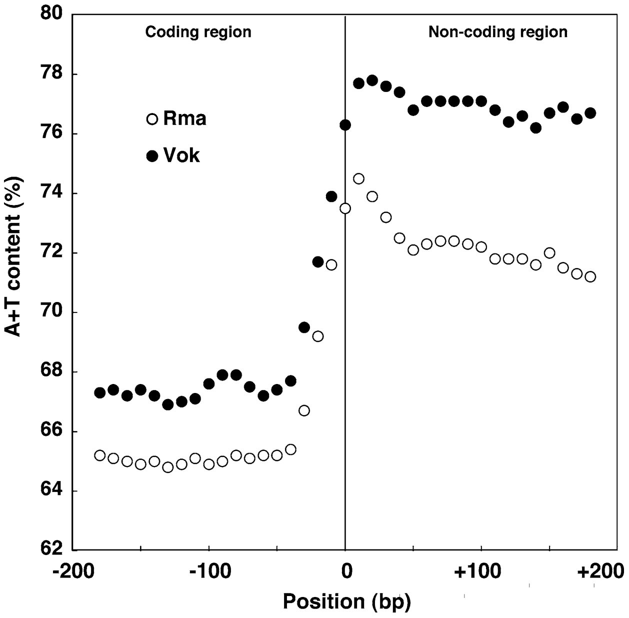


Fig. 6 Profiles of averaged A+T content at the transition of coding and non-coding regions in *Calyptogena okutanii* symbiont (*Vesicomyosocius okutanii*: Vok) and *C. magnifica* symbiont (*Ruthia magnifica*: Rma), respectively. Each *dot* indicates the average A+T content of 402 and 446 transition regions in Vok and Rma, respectively. The A+T content was calculated in windows of 40 bp-DNA, which was shifted by 10 bp from −200 to +200 bp of the transition (*0* on the abscissa) between protein coding and non-coding regions. A+T contents in non-coding regions are significantly higher than those of coding regions

repair genes in these symbionts (Table 2) probably have accelerated mutation rates, and small effective population sizes during vertical transmission have increased the fixation rates of slightly deleterious mutations. However, the lower *dS* value of *Calyptogena* symbionts (0.46 ± 0.15) compared to other symbiont pairs (*B. aphidicola* strains and *Blochmannia* species: 0.83 ± 0.20 and 0.99 ± 0.27) suggests that after divergence *Calyptogena* symbionts have accumulated lesser mutations than the other pairs. These data together with the presence of many large sized deletions (Fig. 2) may imply that *Calyptogena* symbionts are in an earlier stage of RGE than other symbionts.

Fates of functional genes in *Calyptogena* symbiont genomes

While the two genomes share 857 orthologous genes, Vok has 80 specific genes that have no corresponding orthologs in the Rma genome, and Rma has 119 specific genes that are not found in the Vok genome (ESM, Tables S1, S2). It is debatable as to whether this asymmetric genome reduction in the two *Calyptogena* symbionts is the result of differences of adaptation to their environments. For example, Vok has four genes for nitrate respiratory reductase subunits (*narGHIJ*) that are absent from the Rma genome. In natural habitats, *C. okutanii* is half-buried in anoxic or suboxic sediment while *C. magnifica* lives on basalt rocks (Boss and Turner 1980) where a higher oxygen concentration is expected. Because *C. magnifica* may be exposed to oxic seawater, *narGHIJ* are likely unnecessary in Rma.

The Rma genome has a larger number of specific genes in the functional category of “cell wall/membrane” than Vok (ESM, Fig. S2). Rma may retain some functions related to outer and inner membranes which have been lost from Vok. However, no other significant difference was found between the number of Rma and Vok-specific genes in any other particular gene functional categories (ESM, Fig. S2). Most of Rma and Vok-specific genes are probably non-essential for intracellular lifestyle and may be fated to be lost from the genomes in the future. These data indicate that in *Calyptogena* symbionts RGE is occurring, and that the gene repertoires in the two lineages are becoming more similar over time.

Table 3 Rates of nonsynonymous and synonymous substitutions of *Calyptogena* symbionts (Vok–Rma), *Buchnera aphidicola* strains (Buc–Bas), *Blochmannia* species (Bfl–Bpn), and *Escherichia coli–Salmonella typhimurium* (Eco–Stm)

	dN	*dS*	*dN/dS*
Vok–Rma	0.09 ± 0.05	0.46 ± 0.15	0.20 ± 0.09
Buc–Bas	0.15 ± 0.07	0.83 ± 0.20	0.19 ± 0.11
Bfl–Bpn	0.22 ± 0.10	0.99 ± 0.27	0.25 ± 0.17
Eco–Stm	0.03 ± 0.02	0.58 ± 0.37	0.05 ± 0.08

Comparison of 317 orthologs among these organisms were performed

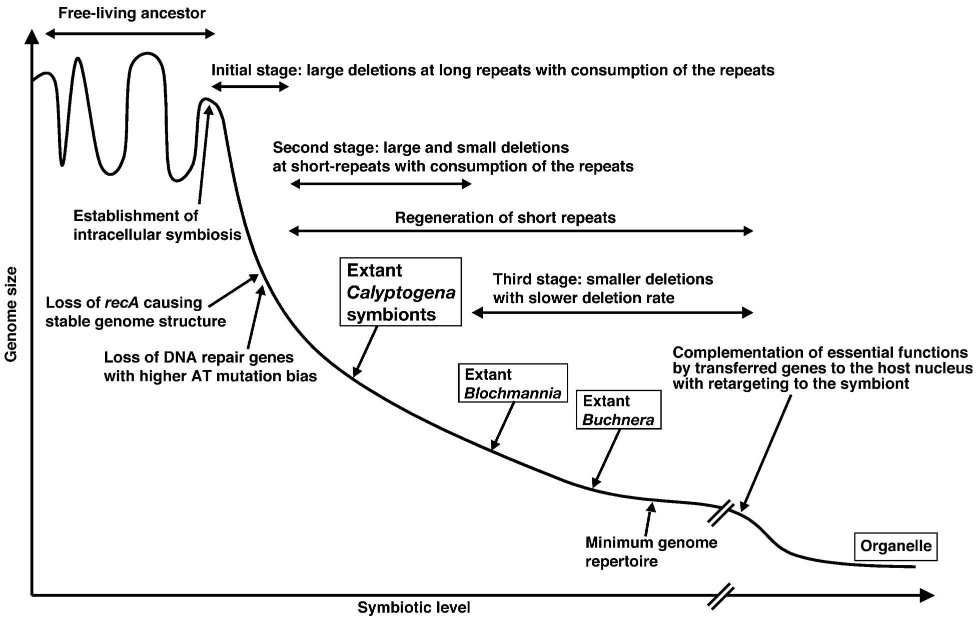
dN Average nonsynonymous base substitution/site, *dS* Average synonymous base substitution/site

Process of reductive genome evolution (RGE) in intracellular symbionts

From the present data, we propose a general scenario for RGEs of *Calyptogena* symbionts and some insect symbionts such as *Buchnera* that have lost *recA* in the process of RGE (Fig. 7).

The ancestor of the symbionts possibly has a larger genome with more structural RNA (tRNA and rRNA) genes and more long (>200 bp) repeats. RGE begins after the establishment of intracellular and vertical transmission type symbiosis with the host. In this initial stage, the symbiont probably still has *recA* and the genome structure

Fig. 7 Proposed process of general reductive genome evolution in intracellular symbionts


Free-living ancestor
Initial stage: large deletions at long repeats with consumption of the repeats
Second stage: large and small deletions at short-repeats with consumption of the repeats
Regeneration of short repeats
Establishment of intracellular symbiosis
Extant *Calyptogena* symbionts
Third stage: smaller deletions with slower deletion rate
Loss of *recA* causing stable genome structure
Loss of DNA repair genes with higher AT mutation bias
Extant *Blochmannia*
Extant *Buchnera*
Complementation of essential functions by transferred genes to the host nucleus with retargeting to the symbiont
Minimum genome repertoire
Organelle
Genome size
Symbiotic level

may change frequently by RecA-dependent homologous recombination. For such recombination, the structural RNA genes and long-repeats are probably important. By the end of this stage, many non-essential genes for the intracellular lifestyle and for recombination and repair including *recA* are likely lost. Most of the long repeats may have been consumed before the loss of *recA*. After the loss of *recA* and other genes involved in DNA recombination and repair, genome structure is likely stabilized but the mutation rate probably increases.

In the next stage, RGE proceeds by a RecA-independent mechanism associated with shorter repeats (<200 bp). Both small and large size deletions (from <100 to over 1 kbp) probably occur in this stage. Non-essential genes may be mutated to non-coding sequences, which are under less functional constraints. While short-repeats are consumed by deletions, the increased mutations with a higher A+T bias probably generates repeated sequences. Subsequently, non-coding sequences are deleted by RecA-independent but short-repeat dependent recombination, such as DRS. The extant *Calyptogena* symbiont genomes are probably at this actively reducing stage.

In the next stage, the third stage, the RGE rate of the symbiont probably slows and reaches a steady state, because the genome is approaching the minimum gene set that is essential for an intracellular symbiotic lifestyle. While large size deletions cease to occur, small size deletions (<100 bp) are the major component of RGE in this stage. Most of the non-essential genes for intracellular symbiotic lifestyle with vertical transmission are likely to be lost by the end of this stage. It seems that the extant *Blochmannia* species are in an early phase of this stage, and the extant *B. aphidicola* strains are in a late phase of this stage.

In other cases, when the host has two distinct symbionts, genomes of symbionts may complement metabolic functions of essential gene products with each other. In such a case, during RGE of one symbiont even essential metabolic genes may be lost. For example, *B. aphidicola* BCc, which co-symbioses with a secondary symbiont in the host cell, has a genome approximately 200 kb smaller (416 kb) than *B. aphidicola* strains APS and Sg (Pérez-Brocal et al. 2006).

In a far later stage, the RGE of symbionts may proceed to an extreme if the functions of the essential genes are complemented by a transfer of genes from symbionts to the hosts nucleus with a re-targeting of the gene products to symbionts as in the case of organelles (Dyall et al. 2004). *Carsonella ruddii*, an endosymbiont of *Pachypsylla venusta* (hackberry petiole gall psyllid), has an extremely reduced genome (160 kb) and probably lacks essential genes for an intracellular symbiotic lifestyle (Nakabachi et al. 2006). Gene transfer from the ancestral symbiont to the host ancestor has been suggested (Nakabachi et al. 2006).

In *Calyptogena* symbionts, RGE is currently ongoing towards smaller size genomes, and at some extreme endpoint these symbionts may eventually become chemoautotrophic organelles.

Acknowledgment Drs. T. Okutani, Y. Fujiwara and K. Fujikura are acknowledged for the discussion about lifestyles of *Calyptogena* clams.

References

Achaz G, Boyer F, Rocha EP, Viari A, Coissac E (2007) Repseek, a tool to retrieve approximate repeats from large DNA sequences. Bioinformatics 23:119–121

Alexeyenko A, Tamas I, Liu G, Sonnhammer EL (2006) Automatic clustering of orthologs and inparalogs shared by multiple proteomes. Bioinformatics 22:e9–e15

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402

Au KG, Clark S, Miller JH, Modrich P (1989) *Escherichia coli mutY* gene encodes an adenine glycosylase active on G-A mispairs. Proc Natl Acad Sci USA 86:8877–8881

Boss KJ, Turner RD (1980) The giant white clam from the Garapagos Rift, *Calyptogena magnifica* species novum. Malacologia 20:161–194

Brudno M, Do CB, Cooper GM, Kim MF, Davydov E, NISC Comparative Sequencing Program, Green ED, Sidow A, Batzoglou S (2003) LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. Genome Res 13:721–731

Bzymek M, Lovett ST (2001) Instability of repetitive DNA sequences: the role of replication in multiple mechanisms. Proc Natl Acad Sci USA 98:8319–8325

Dale C, Wang B, Moran N, Ochman H (2003) Loss of DNA recombinational repair enzymes in the initial stages of genome degeneration. Mol Biol Evol 20:1188–1194

Degnan PH, Lazarus AB, Wernegreen JJ (2005) Genome sequence of *Blochmannia pennsylvanicus* indicates parallel evolutionary trends among bacterial mutualists of insects. Genome Res 15:1023–1033

Delmotte F, Rispe C, Schaber J, Silva FJ, Moya A (2006) Tempo and mode of early gene loss in endosymbiotic bacteria from insects. BMC Evol Biol 6:56

Dyall SD, Brown MT, Johnson PJ (2004) Ancient invasions: from endosymbionts to organelles. Science 304:253–257

Endow K, Ohta S (1990) Occurrence of bacteria in the primary oocytes of vesicomyid clam *Calyptogena soyoae*. Mar Ecol Prog Ser 64:309–311

Frank AC, Amiri H, Andersson SG (2002) Genome deterioration: loss of repeated sequences and accumulation of junk DNA. Genetica 115:1–12

Gil R, Silva FJ, Zientz E, et al (2003) The genome sequence of *Blochmannia floridanus*: comparative analysis of reduced genomes. Proc Natl Acad Sci USA 100:9388–9393

Hall BG (2004) CodonAlign 2.0 published at: http://homepage.mac.com/barryghall/CodonAlign.html

Hanada K, Iwasaki M, Ihashi S, Ikeda H (2000) UvrA and UvrB suppress illegitimate recombination: synergistic action with RecQ helicase. Proc Natl Acad Sci USA 97:5989–5994

Iyer L M, Koonin EV, Aravind L (2002) Classification and evolutionary history of the single-strand annealing proteins, RecT, Redbeta, ERF and RAD52. BMC Genomics 3:8

Klasson L, Andersson SG (2004) Evolution of minimal-gene-sets in host-dependent bacteria. Trends Microbiol 12:37–43

Kuwahara H, Yoshida T, Takaki Y, et al (2007) Reduced genome of the thioautotrophic intracellular symbiont in a deep-sea clam, *Calyptogena okutanii*. Curr Biol 17:881–886

Li WH (1993) Unbiased estimation of the rates of synonymous and nonsynonymous substitution. J Mol Evol 36:96–99

Lovett ST (2004) Encoded errors: mutations and rearrangements mediated by misalignment at repetitive DNA sequences. Mol Microbiol 52:1243–1253

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

Moran NA (2003) Tracing the evolution of gene loss in obligate bacterial symbionts. Curr Opin Microbiol 6:512–518

Moran NA, Mira A (2001) The process of genome shrinkage in the obligate symbiont *Buchnera aphidicola*. Genome Biol 2:Research0054.1–12

Nakabachi A, Yamashita A, Toh H, et al (2006) The 160-kilobase genome of the bacterial endosymbiont *Carsonella*. Science 314:267

Newton IL, Woyke T, Auchtung TA, et al (2007) The *Calyptogena magnifica* chemoautotrophic symbiont genome. Science 315: 998–1000

Nghiem Y, Cabrera M, Cupples CG, Miller JH (1988) The *mutY* gene: a mutator locus in *Escherichia coli* that generates G.C–T.A transversions. Proc Natl Acad Sci USA 85:2709–2713

O'Brien KP, Remm M, Sonnhammer EL (2005) Inparanoid: a comprehensive database of eukaryotic orthologs. Nucleic Acids Res 33:D476–D80

Pamilo P, Bianchi NO (1993) Evolution of the *Zfx* and *Zfy* genes: rates and interdependence between the genes. Mol Biol Evol 10:271–281

Peek AS, Feldman RA, Lutz RA, Vrijenhoek RC (1998) Cospeciation of chemoautotrophic bacteria and deep sea clams. Proc Natl Acad Sci USA 95:9962–9966

Pérez-Brocal V, Gil R, Ramos S, et al (2006) A small microbial genome: the end of a long symbiotic relationship? Science 314: 312–313

Rocap G, Larimer FW, Lamerdin J, et al (2003) Genome divergence in two *Prochlorococcus* ecotypes reflects oceanic niche differentiation. Nature 424:1042–1047

Rocha EP (2003) An appraisal of the potential for illegitimate recombination in bacterial genomes and its consequences: from duplications to genome reduction. Genome Res 13:1123–1132

Tamas I, Klasson L, Canback B, Naslund AK, Eriksson AS, Wernegreen JJ, Sandstrom JP, Moran NA, Andersson SG (2002) 50 million years of genomic stasis in endosymbiotic bacteria. Science 296:2376–2379

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

Tzeng YH, Pan R, Li WH (2004) Comparison of three methods for estimating rates of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol 21:2290–2298

Zhang Z, Li J, Zhao XQ, Wang J, Wong GK, Yu J (2006) KaKs_Calculator: calculating Ka and Ks through model selection and model averaging. Genomics Proteomics Bioinformatics 4:259–263