Genome size reduction through multiple events of gene disintegration in *Buchnera APS*

Francisco J. Silva, Amparo Latorre and Andrés Moya

The evolution of the endosymbiont *Buchnera* during its adaptation to intracellular life involved a massive reduction in its genome. By comparing the orthologous genes of *Buchnera, Escherichia coli* and *Vibrio cholerae*, we show that the minimal genome size of *Buchnera* arose from multiple events of gene disintegration dispersed over the whole genome. The elimination of the genes was a continuous process that began with gene inactivation and progressed until the DNA corresponding to the pseudogenes were completely deleted.

Many intracellular bacterial species, either pathogenic or mutualist, have small size genomes compared with their free-living relatives. The reduction of genome size is associated with the loss of a great number of genes as an adaptation to new conditions where many molecules can be obtained from the host and do not need to be synthesized. However, in several cases not only the genes, but also the derived pseudogenes have disappeared, leading to a highly compact genome. Several mechanisms based on serial deletions or chromosomal rearrangements have been proposed to explain genome-size decline. However, a detailed study focused on the reduction of the complete genome of an intracellular bacterium has never been reported.

Here, we have applied a whole-genome approach to understand the genome reduction of the bacterium *Buchnera* sp. APS, the primary endosymbiont of aphids. The *Buchnera* genome has recently been sequenced and contains a 641-kb chromosome with 600 genes. Our results reveal that a mechanism of gene-by-gene disintegration, affecting multiple locations in the genome, was important during this reductive evolution.

First, we compared the gene contents and gene locations between *Buchnera* and *Escherichia coli* K-12 (which has a 4.6-Mb chromosome with 4404 genes), the closest free-living relative of *Buchnera* with a completely sequenced genome. We detected segments containing genes that maintain synteny, despite the many chromosomal rearrangements that have occurred since the divergence of the species, ~200 Myr ago. Examining genes of both species, we found that 583 out of the 600 *Buchnera* genes have unquestionably orthologs in *E. coli*, including structural RNAs and coding genes. We identified candidate *E. coli* (and *V. cholerae*, see below) genes that have an ortholog in *Buchnera* using the Microbial Genome Database for Comparative Analysis (http://mbgd.genome.ad.jp) and BLAST, with a cutoff value of 0.001 and a minimum overlapping region of 30% due to the smaller size of several *Buchnera* genes. We tested the orthology of some genes in detail by comparing size and similarity, and performing phyletogenetic reconstructions with members of the Proteobacteria.

Then, we aligned the genomes (Fig. 1a) using the chromosomal positions of the genes and taking the origin of replication as bp 0 in *E. coli*. Each point in Fig. 1a represents the position of the orthologous gene pair in the *Buchnera* (x-axis) and *E. coli* (y-axis) chromosome. We observed that many chromosomal rearrangements have occurred since the divergence of the species, the most frequent ones being the inversions around the origin and terminus of replication, generating the pattern recently described as X alignment.  

This comparison found 496 orthologous genes among the three genomes, we looked for these ancestral DNA segments that we will refer to as 'blocks'. A block is a conserved gene order.

![Fig. 1. Genomic alignment of orthologous gene position. Plots show the chromosome locations of pairs of orthologous genes. (a) *E. coli* versus *Buchnera* APS. (b) *V. cholerae* versus *Buchnera* APS. Red lines mark the regions around replication termini, which are the chromosome regions opposite to the origin of replication. They might contain specific regulatory sequences such as the tarsites in *E. coli*.](http://tig.trends.com/0168-9525/01/5--see-front-matter©2001ElsevierScienceLtd.Allrightsreserved.PII:S0168-9525(01)02483-0)
segment of DNA formed by two or more genes that must fulfill the following characteristics:
(1) The two genes placed at each end are the same in the three species.
(2) Every gene of the block must be present in E. coli and V. cholerae. Those genes contained in only one of the two reference species (E. coli and V. cholerae) are not accepted as part of the block, because they are considered putative lateral gene transfer insertions or false genes.
(3) The transcriptional orientation of each gene, starting from the same end, must be the same in the three species.
(4) A gene absent in the Buchnera block must not be present elsewhere in the Buchnera chromosome.

Using this restrictive approach, we identified 95 ancestral blocks of genes (containing 502 genes). The number of genes of each block is variable, the largest being formed by 28 genes, corresponding mainly to ribosomal protein genes. Of the ancestral blocks, 49 have no genes lost in Buchnera; however, one or several genes are absent from 46 blocks (111 genes in total, including five pseudogenes). The distribution of these absent genes (Table 1) shows a considerable number of single genes that are absent (33 genes).

Table 1. Distribution of genes absent from the ancestral blocks in Buchnera sp. APS

<table>
<thead>
<tr>
<th></th>
<th>Single gene loss</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Two</td>
<td>Three</td>
<td>Four</td>
<td>Five</td>
<td>Six</td>
<td></td>
</tr>
<tr>
<td>No. of cases</td>
<td>33</td>
<td>15</td>
<td>7</td>
<td>4</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>No. of genes</td>
<td>33</td>
<td>30</td>
<td>21</td>
<td>16</td>
<td>5</td>
<td>6</td>
</tr>
</tbody>
</table>

In other cases, two or more adjacent genes have disappeared, with the largest number of adjacent genes lost being six. Figure 2 shows the putative ancestral genome content of E. coli and Buchnera. It follows the gene order of E. coli K-12 strain. Genes with no ortholog in either V. cholerae or Buchnera were not included. The minimal ancestral genome (Fig. 2) contains 1818 genes, but it does not include most of the structural RNAs genes, due to the difficulty in assigning orthology to them. Outside the blocks, we observe stretches of up to 48 absent ancestral genes in Buchnera, but because they do not form part of the blocks, they were probably not in this order in the ancestor of Buchnera, and therefore it is improbable that they were lost in a same deletion event.

By analyzing the intergenic regions of Buchnera that correspond to regions containing one or several genes in the ancestor’s genome, we have been able to detect cases corresponding to the three stages of gene decay (Fig. 3) that support a model of gene-by-gene disintegration. The first stage is where the gene is inactivated (pseudogene) but almost all of its sequence remains (e.g. fabD pseudogene, Fig. 3a). In the second stage, the DNA sequence that corresponded to the gene is gradually denuded by continual mutation and small deletions, leaving only a part of the ancestral sequence that is no longer recognizable as a pseudogene owing to the strong A+T mutational bias of these bacterial endosymbionts (e.g. dam and damX genes, Fig. 3b). In the final stage, the ancestral sequence has been lost entirely (e.g. apaG and pdxA genes, Fig. 3c).

To examine distribution of the three stages of gene decay, we analyzed the intergenic regions of the Buchnera genome, finding mean and median sizes of 116.4 and 74 bp, respectively. These are slightly smaller than the sizes of the intergenic regions of the remnant sequence of the lost genes (157.1 and 125). Based on these values, we arbitrarily set 200 bp as the limit between the second and third stages of decay (Table 2). The quantification of the size of the intergenic regions where genes are absent shows that the reductive forces that acted during the evolution of the Buchnera genome were so strong that in 96 out of 111 cases the DNA sequence of the inactivated genes has been completely or nearly completely removed. We detected only ten genes where the DNA sequence has not been completely removed and five pseudogenes that are probably the result of recent inactivations, because they have a G+C content similar to the active genes in the Buchnera genome.

Fig. 2. The genome of the common ancestor of E. coli and Buchnera. Each circle represents a gene. Red, genes conserved in Buchnera; blue, genes not present in the Buchnera genome; orange, pseudogenes in Buchnera. Genes are ordered according to the E. coli K-12 genome, and the first circle corresponds to E. coli b0002 (thrA). The origin of replication is marked with a black circle. Ancestral ‘blocks’ (see text) are shown as green rectangles. This genome includes those E. coli genes with an ortholog in either V. cholerae or Buchnera APS.

To the gene is gradually denuded by continual mutation and small deletions, leaving only a part of the ancestral sequence that is no longer recognizable as a pseudogene owing to the strong A+T mutational bias of these bacterial endosymbionts (e.g. dam and damX genes, Fig. 3b). In the final stage, the ancestral sequence has been lost entirely (e.g. apaG and pdxA genes, Fig. 3c).

To examine distribution of the three stages of gene decay, we analyzed the intergenic regions of the Buchnera genome, finding mean and median sizes of 116.4 and 74 bp, respectively. These are slightly smaller than the sizes of the intergenic regions of the remnant sequence of the lost genes (157.1 and 125). Based on these values, we arbitrarily set 200 bp as the limit between the second and third stages of decay (Table 2). The quantification of the size of the intergenic regions where genes are absent shows that the reductive forces that acted during the evolution of the Buchnera genome were so strong that in 96 out of 111 cases the DNA sequence of the inactivated genes has been completely or nearly completely removed. We detected only ten genes where the DNA sequence has not been completely removed and five pseudogenes that are probably the result of recent inactivations, because they have a G+C content similar to the active genes in the Buchnera genome.

Similar mechanisms, in combination with homologous recombination, seem to have produced the single unlinked 235 + 5S and 16S rRNA genes in Buchnera from the putative seven or eight complete clusters of the ancestor (Fig. 4). The simplest explanation for the origin of the Buchnera 235 + 5S + tRNA_Glu cluster (Fig. 4b) is a recombination event between the 16S rRNA gene of the rrn cluster adjacent to the aroE and yrdC genes (equivalent position to E. coli rrnD and V. cholerae rrnA clusters) with a ribosomal cluster in opposite orientation (rrnC, B or E in E. coli) with the consequent replacement of two tRNAs (Ala and Ile) by the tRNA_Glu, followed by the decay of the 16S rRNA gene and two coding genes. The unlinked 16S rRNA cluster has a similar position to the E. coli rrnH, based on the downstream presence of orthologous genes and pseudogenes (gloB, vrrmA and dnaQ). These genes are downstream of an rrn cluster in several Salmonella species and in Yersinia pestis (EnteriX
server, http://galapagos.cse.psu.edu/enterix), but not in V. cholerae. This implies the specific loss of the 23S rRNA, 5S and a tRNAAsp and some coding genes in the Buchnera cluster (Fig. 4c).

The process of gene loss during adaptation of Buchnera to intracellular life mainly involved multiple, dispersed gene inactivation events. There is only one large region around the E. coli terminus of replication (Fig. 1a) with very few orthologs in Buchnera. The different gene composition of this region between two E. coli strains and the similar lack of orthology between the E. coli and V. cholerae termini, indicates that this region is specially unstable and dispensable, making impossible to predict its gene content in the common ancestor. In fact, this region is not present in our proposed ancestral genome.

In this comparative study, we analyzed a significant sample of genes (502), which corresponds to ~25% of our proposed ancestral genome content. Although the selection of the analyzed genes is not random, being based on the criterion of inclusion in a block, we believe that the conclusions obtained from their analysis can be extended to the rest of the genome. We propose that small-scale changes in the genome would be able to cause gene disintegration at multiple sites and generate the reduced genome of this bacterium. However, other mechanisms, such as large deletions, cannot be ruled out. In support of our proposal, around a third of the genes absent from the Buchnera genome (Table 1) were lost individually, rather than as a larger deletion of several adjacent genes. However, this individual gene loss can be extended to other analyzed genes. For example, two out of the 15 losses of two adjacent genes correspond to the absence of a gene plus the presence of a pseudogene. We propose that, with time, the DNA sequence of the pseudogene would disappear, causing the loss of two adjacent genes that would be erroneously interpreted as a large deletion of the two genes. Thus, the distribution of absent genes by intergenic region (Table 1) need not be explained by large deletional events of adjacent genes. On the contrary, an elemental rule of probability explains how a mechanism affecting single genes can produce, with decreasing probabilities, absences of two, three or more adjacent genes.

The mechanism of gene disintegration presented in this paper has been described recently in several Rickettsia species, where a detailed study of pseudogenes has shown a more advanced stage of degradation that the authors called fossil open reading frames. The most significant thing was that, although some larger deletions were observed (599 and 767 bp), the mean and median size of the deletions were small (51.2 and 4 bp, respectively). These values are in agreement with the idea of gene disintegration by a continuous mechanism of small deletions that decrease the size of the pseudogenes. However, the Buchnera genome is an extreme when compared with other intracellular bacteria having larger genomes and larger amounts of noncoding DNA, such as Rickettsia or Mycobacterium leprae. In Buchnera, most of the genes lost from the ancestral blocks are in an advanced stage of disintegration – the DNA from 96 out of 111 lost genes has practically disappeared, and only a few genes are in an intermediate stage (e.g. 884 bp in the intergenic region where dam and damX genes were located, see Fig. 3b).

The accumulation of deleterious mutations in Buchnera caused by continual bottlenecks owing to its maternal inheritance in aphid lineages, as well as by selection favoring Buchnera shorter genomes within the bacteriocyte could explain the reduction of the genome and the extremely low number of extant pseudogenes. Selection for shorter genomes might occur at two levels. First, within each polyploid bacterium, there will be competition among genomic molecules, and the shortest will probably be fixed in the bacterium, owing to faster replication. Second, as long as they are not deleterious, smaller genomes can accelerate bacterium duplication, thus favoring survival of the bacterium. These mechanisms have already been proposed for reductive evolution in endosymbiotic organelles. The production of small deletions would be favored by the bias towards A+T content, probably increased as a consequence of the loss of repair genes. Replication slippage or intramolecular recombination between A+T-rich direct repeats would increase the

---

**Table 2. Stages of gene disintegration in ancestral blocks from Buchnera sp. APS**

<table>
<thead>
<tr>
<th>Pseudogenes</th>
<th>Intergenic regions (&gt;200 bp)</th>
<th>Intergenic regions (&lt;200 bp)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of genes</td>
<td>5</td>
<td>10</td>
</tr>
</tbody>
</table>

*In those intergenic regions where more than one gene is absent, the size (bp) is divided by the number of genes.*
deletion rate. Genetic drift linked to a high deletion rate in Buchnera might help the process of genome reduction.

Acknowledgements
This work was supported by grant BFM2000-1383 from the Ministerio de Ciencia y Tecnología (Spain). We thank I. Marín, R. Gil and F. González-Candelas for critical review of the manuscript.

References

Francisco J. Silva*
Amparo Latomé
Andrés Moya
Institut Cavanilles de Biodiversitat i Biologia Evolutiva and Departament de Genética, Universitat de València, Apartado 22085, 46071 València, Spain.
*e-mail: Francisco.Silva@uv.es

Editorial policy for Genome Analysis papers
The Trends in Genetics Genome Analysis section includes original observations concerning the function, organization and evolution of genomes. All sequences used in the analysis must be available through GenBank. Although other articles in TIG are generally invited by the Editor, authors wishing to publish a Genome Analysis paper should send a synopsis of their proposed article directly to the Advisory Editor Dr E.V. Koonin at National Center for Biotechnology, NIH, Bethesda, MD 20894, USA (e-mail: koonin@ncbi.nlm.nih.gov).

For further information, please e-mail: tig@current-trends.com

http://tig.trends.com