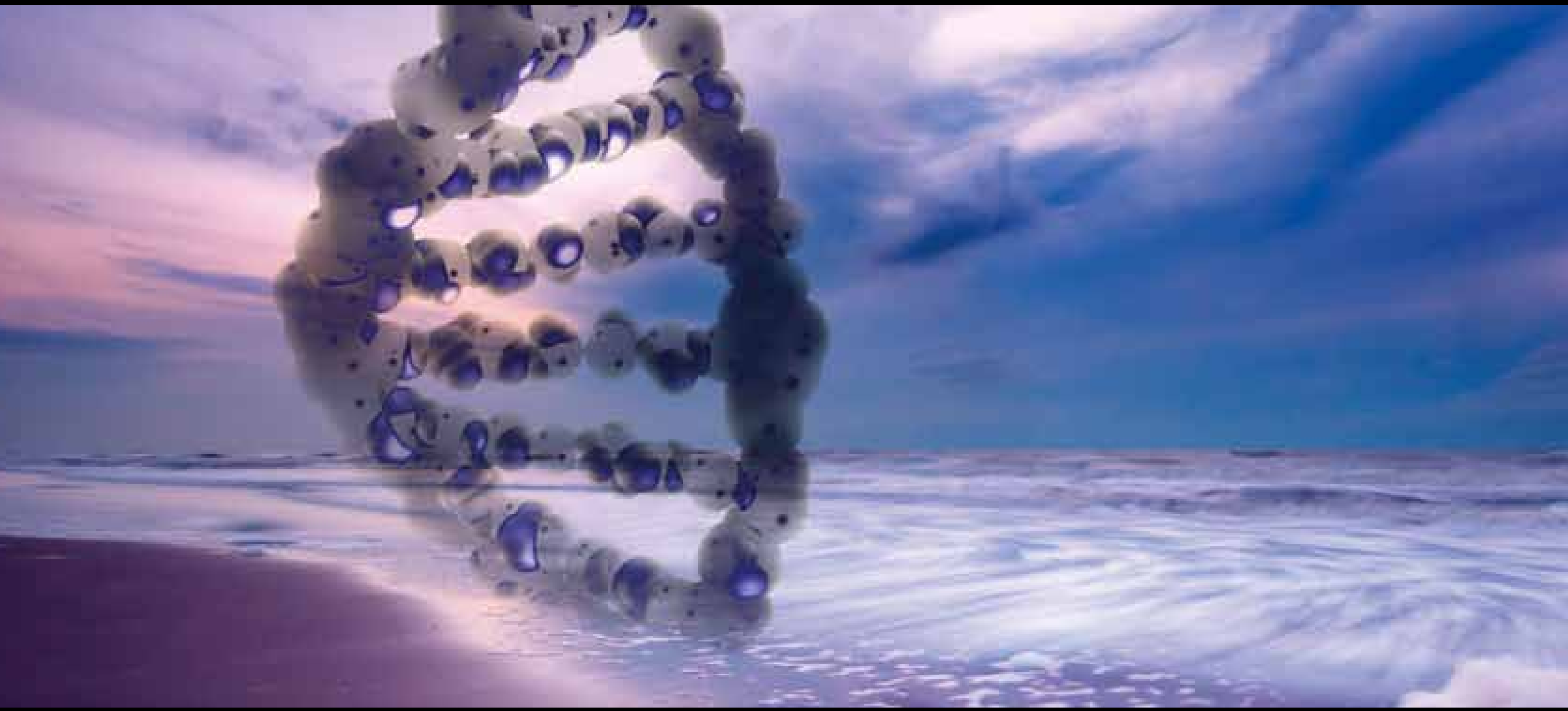
International Journal of Evolutionary Biology

Evolutionary Mechanisms of Microbial Genomes

Guest Editors: Hiromi Nishida, Shinji Kondo, Hideaki Nojiri, Ken-ichi Noma, and Kenro Oshima



Evolutionary Mechanisms of Microbial Genomes

International Journal of Evolutionary Biology

Evolutionary Mechanisms of Microbial Genomes

Guest Editors: Hiromi Nishida, Shinji Kondo,
Hideaki Nojiri, Ken-ichi Noma, and Kenro Oshima

International Journal of Evolutionary Biology

Editorial Board

Contents

SAGE-Hindawi Access to Research
International Journal of Evolutionary Biology
Volume 2011, Article ID 319479, 2 pages
doi:10.4061/2011/319479

Editorial

Evolutionary Mechanisms of Microbial Genomes

Hiromi Nishida,[1] Shinji Kondo,[2] Hideaki Nojiri,[3] Ken-ichi Noma,[4] and Kenro Oshima[5]

[1] *Agricultural Bioinformatics Research Unit, Graduate School of Agricultural and Life Sciences, The University of Tokyo, Tokyo 113-8657, Japan*
[2] *Laboratory for Cellular Systems Modeling, RIKEN Research Center for Allergy and Immunology, Kanagawa 230-0045, Japan*
[3] *Laboratory of Environmental Biochemistry, Biotechnology Research Center, The University of Tokyo, Tokyo 113-8657, Japan*
[4] *Gene Expression and Regulation Program, The Wistar Institute, Philadelphia, PA 19104, USA*
[5] *Department of Agricultural and Environmental Biology, Graduate School of Agricultural and Life Sciences, The University of Tokyo, Tokyo 113-8657, Japan*

Correspondence should be addressed to Hiromi Nishida, hnishida@iu.a.u-tokyo.ac.jp

Received 3 April 2011; Accepted 3 April 2011

Sequencing of more than 1,600 microbial genomes has been complete, and rigorous studies are underway to reveal the mechanisms of evolution which gave rise to the great variety in combination of gene functions encoded on these genomes. Although comparative genomics based on orthologous genes has elucidated a great deal of the phylogenetic relationship among the sequenced genomes, the mechanisms which have shaped the current states of the microbial genomes remain elusive. Particularly, the contribution of external forces such as horizontal gene transfer and pressure from environmental factors to genome evolution has yet to be investigated. This special issue presents six, three, and two papers related, respectively, to bacterial, fungal, and viral evolutionary mechanisms.

Among the six papers regarding bacterial evolution, H. Nishida and C. -S. Yun in "*Phylogenetic and guanine-cytosine content analysis of* Symbiobacterium thermophilum *genes*" reported a mechanism of the *Symbiobacterium* genome which increased GC content of horizontally transferred genes and thereby maintained the genome with high GC content. K. Oshima et al. in "*Unique evolution of* Symbiobacterium thermophilum *suggested from gene content and orthologous protein sequence comparisons*" performed phylogenetic analyses of more than 50 Clostridia by comparing gene content and orthologous protein sequence and demonstrated that these two phylogenetic relationships are topologically different, strongly suggesting that each Clostridia has a species-specific gene content likely due to frequent genetic exchanges or gene losses which have occurred during evolution. H. -Y. Dou et al. in "*Prevalence of* Mycobacterium tuberculosis *in Taiwan: a model for strain evolution linked to population migration*" presented an association study of distinct *Mycobacterium tuberculosis* strains prevalent in Taiwan with historical migrations of different ethnic populations based on a comparison of the tandem repeat sequences as genetic markers. T. Takeda et al. in "*Distribution of genes encoding nucleoid-associated protein homologs in plasmids*" reported biases associated with certain bacterial plasmids, that is, increase of nucleoid-associated protein genes in large bacterial plasmids and low GC content of plasmids encoding (histone-like nucleoid structuring protein) H-NS. V. Pérez-Brocal et al. in "*New insights on the evolutionary history of aphids and their primary endosymbiont* Buchnera aphidicola" presented a study which supports the hypotheses of divergence of *Buchnera aphidicola* from their host lineages during an early Cretaceous period by demonstrating a closer relationship of a subfamily Eriosomatinae with Lachninae than with Aphidinae. A. Moreno-Letelier et al. in "*Parallel evolution and horizontal gene transfer of the* pst *operon in* Firmicutes *from oligotrophic environments*" demonstrated that the phosphate transport system gene operon of Firmicutes has two highly divergent clades which do not correlate either with the type of habitat or with a phylogenetic congruence and proposed parallel evolution of this gene after horizontal gene transfer events.

Of the three papers dealing with fungal evolution, R. T. Morris and G. Drouin in "*Ectopic gene conversions in the genome of ten hemiascomycete yeast species*" found that ectopic gene conversions in the genome of ten hemiascomycetes tend to occur more frequently between closely linked genes and proposed that the mechanisms responsible for the loss of introns in *Saccharomyces cerevisiae* were also involved in the 3′-end gene conversion bias observed among the paralogs. E. van Zijll de Jong et al. in "*Sequence analysis of SSR-flanking regions identifies genome affinities between pasture grass fungal endophyte taxa*" demonstrated that some asexual *Neotyphodium* species arose following interspecies hybridization between sexual *Epichloe* ancestors and characterized *Neotyphodium* isolates based on sequence analysis of genomic regions flanking simple sequence repeats. N. Khaldi and K. H. Wolfe in "*Evolutionary origins of the fumonisin secondary metabolite gene cluster in* Fusarium verticillioides *and* Aspergillus niger" compared the fumonisin secondary met abolite gene cluster and proposed that the gene cluster was horizontally transferred to *Aspergillus niger* from a Sordariomycete.

As for the two papers of viral evolution, K. Tang and X. Wu in "*Computational analysis suggests that Lyssavirus glycoprotein gene plays a minor role in viral adaptation*" found no significant evidence of positive selection on any site of the Lyssavirus glycoprotein-coding gene (except for AY987478) and proposed that the glycoprotein gene has been under purifying selection and that the evolution of this gene may not play a significant role in Lyssavirus adaptation. S. A. B. Miele et al. in "*Baculovirus: molecular insights on their diversity and conservation*" reported an evidence which supports the current division of the Baculoviridae into four genera, *Alpha*-, *Beta*-, *Gamma*-, and *Deltabaculovirus* based on comparative studies of 57 genome sequences from baculoviruses.

In closing this introduction to the special issue, we would like to express our full appreciation to all the authors and reviewers for their enormous efforts that have made the timely completion of our assignment successful. We sincerely hope that this special issue will stimulate further the investigation of evolutionary mechanisms of microbial genomes.

Hiromi Nishida
Shinji Kondo
Hideaki Nojiri
Ken-ichi Noma
Kenro Oshima

SAGE-Hindawi Access to Research
International Journal of Evolutionary Biology
Volume 2011, Article ID 634505, 5 pages
doi:10.4061/2011/634505

Research Article

Phylogenetic and Guanine-Cytosine Content Analysis of *Symbiobacterium thermophilum* Genes

Hiromi Nishida and Choong-Soo Yun

Agricultural Bioinformatics Research Unit, Graduate School of Agriculture and Life Sciences, The University of Tokyo, Bunkyo-ku, Tokyo 113-8657, Japan

Correspondence should be addressed to Hiromi Nishida, hnishida@iu.a.u-tokyo.ac.jp

Received 10 September 2010; Revised 20 October 2010; Accepted 5 November 2010

Academic Editor: Shinji Kondo

Although the bacterium *Symbiobacterium thermophilum* has a genome with a high guanine-cytosine (GC) content (69%), it belongs to a low GC content bacterial group. We detected only 18 low GC content regions with 5 or more consecutive genes whose GC contents were below 65% in the genome of this organism. *S. thermophilum* has 66 transposase genes, which are markers of transposable genetic elements, and 38 (58%) of them were located in the low GC content regions, suggesting that *Symbiobacterium* has a similar gene silencing system as *Salmonella*. The top hit (best match) analyses for each *Symbiobacterium* protein showed that putative horizontally transferred genes and vertically inherited genes are scattered across the genome. Approximately 25% of the 3338 *Symbiobacterium* proteins have the highest similarity with the protein of a phylogenetically distant organism. The putative horizontally transferred genes also have a high GC content, suggesting that *Symbiobacterium* has gained many DNA fragments from phylogenetically distant organisms during the early stage of Firmicutes evolution. After acquiring genes, *Symbiobacterium* increased the GC content of the horizontally transferred genes and thereby maintained a genome with a high GC content.

1. Introduction

Symbiobacterium thermophilum is a syntrophic bacterium that grows effectively when cocultured with a cognate *Geobacillus* sp. [1]. Because of the lack of carbonic anhydrase in the course of *Symbiobacterium* evolution [2, 3], the major growth factor for this organism is CO_2 generated by the growth of *Geobacillus* [4]. *S. thermophilum* has a 3.57 Mbp circular genome that consists of 3338 protein-coding sequences [3]. On the basis of the comparative genomic studies, *Symbiobacterium* is classified as a member of the class Clostridia [3, 5]. Although *Symbiobacterium* phylogenetically belongs to Clostridia (low guanine-cytosine (GC) content bacterial group), the species *S. thermophilum* has a genome with a high GC content (69%).

GC content is commonly used as a marker in bacterial systematics; for example, actinobacteria have a high GC content genome, and clostridia have a low GC content genome. This variation in nucleotide content in bacteria is not clearly understood [6–8]. Analyzing a high GC content genome of a bacterium that belongs to a low GC content group or vice versa is useful and important. *Symbiobacterium* belongs to the class Clostridia (low GC content group), but its genome has a high GC content (69%). One possibility is that *Symbiobacterium* has acquired this high GC content from DNA fragments through horizontal gene transfer [9, 10], homologous gene recombination [11, 12], or both. Another possibility is that *Symbiobacterium* has increased the GC content of the acquired genes and maintained the high GC content during evolution. In this study, we identified GC content of each gene of *S. thermophilum*. In addition, we identified the horizontally transferred and vertically inherited genes. In order to elucidate why the *Symbiobacterium* genome has a high GC content, we compared the GC contents of the horizontally transferred genes with those of the vertically inherited genes.

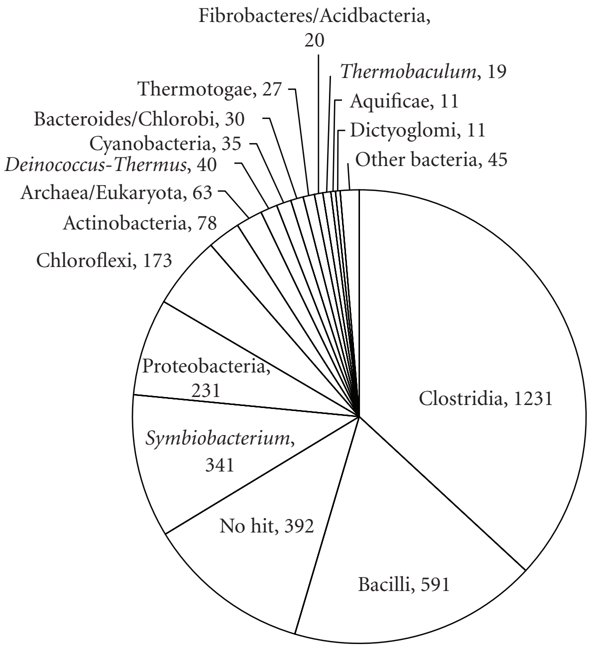


Figure 1: Pie chart of the categories of the 3338 *Symbiobacterium* protein-coding genes. A BLAST search was conducted for all proteins from 147 eukaryotes, 1047 bacteria, and 84 archaea in the KEGG database (http://www.kegg.jp) considering the parameter values given on the GenomeNet website (http://www.genome.jp). The query amino acid sequence was each protein of *Symbiobacterium thermophilum*. The top hit (best match) for each *Symbiobacterium* protein was recorded. However, if the top hit was absent or if the *E*-value of the top hit exceeded 0.1, the *Symbiobacterium* protein was considered to have no similar protein (category, "No hit"). We categorized the 3338 *Symbiobacterium* proteins into the following 17 categories: "Actinobacteria," "Aquificae," "Archaea/Eukaryota," "Bacilli," "Bacteroidetes/Chlorobi," "Chloroflexi," "Clostridia," "Cyanobacteria," "*Deinococcus-Thermus,*" "Dictyoglomi," "Fibrobacteres/Acidobacteria," "No hit," "Proteobacteria," "*Symbiobacterium,*" "*Thermobaculum,*" "Thermotogae," and "Other bacteria." If the top hit was another protein(s) of *Symbiobacterium*, then the query protein was considered to belong to the category "*Symbiobacterium.*"

2. Materials and Methods

In this study, we classified the 3338 protein-coding sequences of *S. thermophilum* on the basis of the amino acid sequence of each coded protein. A BLAST search was conducted for all proteins from 147 eukaryotes, 1047 bacteria, and 84 archaea in the KEGG database (http://www.kegg.jp/) considering the parameter values given on the GenomeNet website (http://www.genome.jp/). The top hit (best match) for each *Symbiobacterium* protein was recorded. However, if the top hit was missing or if the *E*-value of the top hit exceeded 0.1, we considered the *Symbiobacterium* protein to have no similar protein (category, "No hit"). If the top hit was another protein(s) of *Symbiobacterium* (category, "*Symbiobacterium*"), the protein-coding gene was considered to have duplicated during evolution. Top hit analysis at the genome level is a powerful tool for elucidating the phylogenetic lineage of an organism [13, 14].

3. Results and Discussion

On the basis of the phylogenetic lineage of the organism possessing the top hit protein shown in the BLAST result, the 3338 *S. thermophilum* protein-coding genes were classified into 17 categories (Figure 1). The largest category was "Clostridia," and the second largest category was "Bacilli." This is consistent with the results of previous phylogenetic analyses [3]. The third and fourth largest categories were "No hit" and "*Symbiobacterium,*" respectively (Figure 1). Most genes belonging to the category "*Symbiobacterium*" might share their origin with other genes of the same category because 300 of the 341 genes had a similar protein sequence as that of the other organisms that appeared below the top hit of the BLAST result (Table 1 (see Supplementary matrial available online at doi:10.4061/2011/634505)). For example, most transposable elements belonged to "*Symbiobacterium,*" indicating that they were duplicated on the *Symbiobacterium* genome after invasion.

When each gene was plotted on the basis of its category, we detected 52 clusters containing 5 or more consecutive genes belonging to "Clostridia" (Figure 2, pink regions in Supplementary Table 1). These conserved gene clusters are probably not acquired by horizontal gene transfer and are strongly considered to be vertically inherited. The putative vertically inherited genes were scattered across the genome of *S. thermophilum* (Figure 2). In addition, we detected 18 low GC content regions containing 5 or more consecutive genes whose GC contents were below 65% (Figure 3, yellow regions in Supplementary Table 1). These low GC content regions

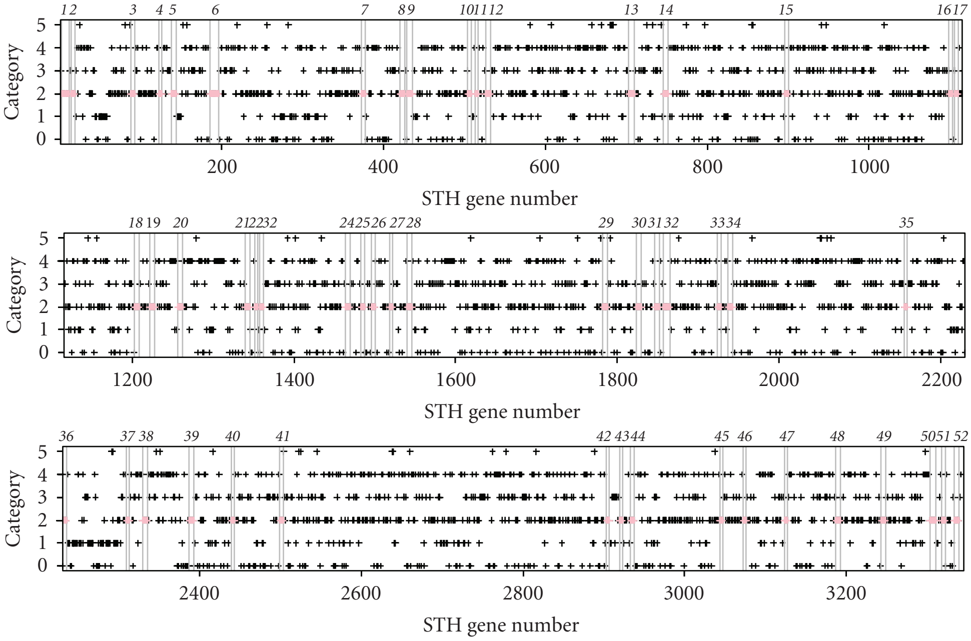


Figure 2: Plots of the location and category of the *Symbiobacterium* protein-coding genes. *X*-axis: STH gene number. *Y*-axis: 0: category "No hit" (*Symbiobacterium*-specific genes); 1: category "*Symbiobacterium*" (multiple copied genes); 2: category "Clostridia;" 3: category "Bacilli"; 4: categories "Actinobacteria," "Aquificae," "Bacteroidetes/Chlorobi," "Chloroflexi," "Cyanobacteria," "*Deinococcus-Thermus*," "Dictyoglomi," "Fibrobacteres/Acidobacteria," "Proteobacteria," "*Thermobaculum*," "Thermotogae," and "Other bacteria"; 5: category "Archaea/Eukaryota." The italicized numbers indicate 52 clusters (pink) containing 5 or more consecutive genes belonging to the category "Clostridia."

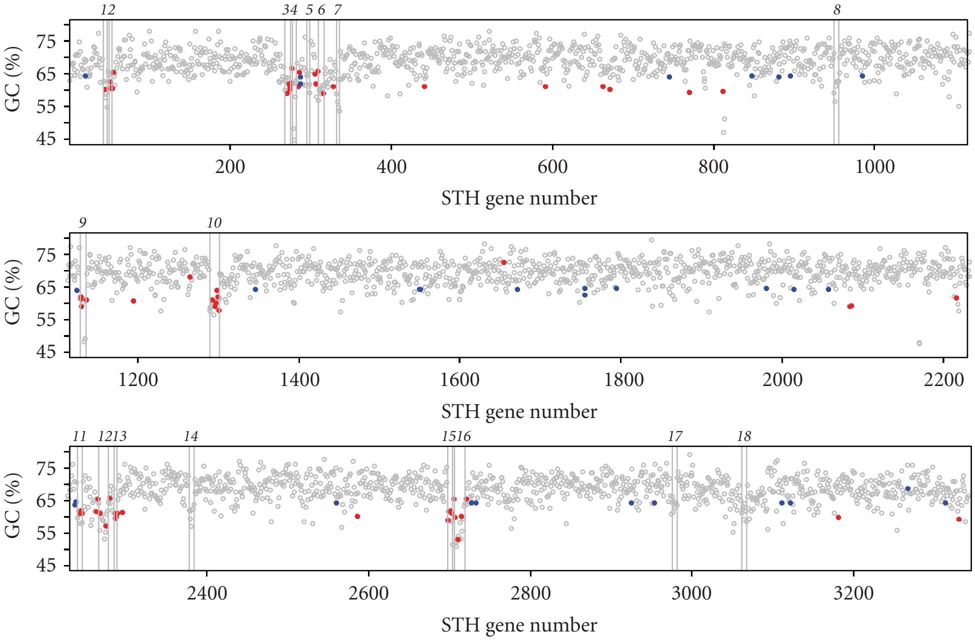


Figure 3: Plots of the location and GC content of the *Symbiobacterium* protein-coding genes. *X*-axis: STH gene number. *Y*-axis: GC content (%) of gene. Red indicates the putative transposase-coding genes, and blue indicates the group II intron-coding maturase genes. The italicized numbers indicate 18 low GC content regions containing 5 or more consecutive genes whose GC contents are below 65%.

do not overlap with the 52 vertically inherited clusters (Supplementary Table 1). On the basis of the KEGG gene cluster database, we found 12 gene clusters in the 18 low GC content regions (Supplementary Table 2).

Approximately 25% of the 3338 *Symbiobacterium* protein-coding genes belonged to categories consisting of organisms phylogenetically distant from *Symbiobacterium* (Figure 1), suggesting that *Symbiobacterium* frequently acquired genes during evolution. The proportion of horizontally transferred genes in the *Symbiobacterium* genome is strongly suggested to be the highest among bacteria [15]. These putative horizontally transferred genes are scattered across the genome of *S. thermophilum* (Figure 2). In addition, considering the species diversification of Bacilli and Clostridia, it is suggested that the categories "Bacilli" and "Clostridia" include not only vertically inherited genes but also horizontally transferred genes.

Transposase genes are generally used as markers of transposable genetic elements [16]. Most transposase-coding genes flank horizontally transferred genes [13]. *S. thermophilum* has 66 putative transposase-coding genes, of which 38 (58%) are located in the low GC content regions (P-value $= 1.3 \times 10^{-99}$; Pearson's chi-square test) (Figure 3), suggesting that *Symbiobacterium* has a similar silencing system as that of *Salmonella* [17, 18]. In the silencing system, a histone-like nucleoid structuring (H-NS) protein binds to the region with a low GC content. Similar functional (H-NS) proteins were reported in *Mycobacterium* and *Pseudomonas* [19, 20]. If *Symbiobacterium* also has such proteins that bind the low GC content regions, the expression of the transposable elements located in these regions might be inhibited. As mentioned above, most transposable elements belong to the category "*Symbiobacterium,*" which is consistent with the fact that the genes of this category have lower GC contents than those of the other categories (Supplementary Figure 1). The regions consisting of low GC content genes cannot be explained by the directional mutation pressure or amelioration of bacterial genomes [21, 22]. Interestingly, although H-NS proteins bind the low GC content regions in *Mycobacterium*, *Pseudomonas*, and *Salmonella* [17–20], the H-NS protein of *Escherichia coli* does not specifically bind only these regions [23].

In addition, *S. thermophilum* has 30 group II intron-encoding maturase genes. Group II introns are transposable elements [24] that encode maturase as an intron-specific splicing factor [25]. The GC content of each maturase gene is approximately 65% (Figure 3). These maturase genes are classified in "*Symbiobacterium,*" on the basis of amino acid sequence similarity. In contrast to the transposase genes, the group II intron-encoding maturase genes are not located in the 18 low GC content regions (Figure 3). If *Symbiobacterium* has both an H-NS protein binding the low GC content regions and a gene silencing system similar to *Mycobacterium*, *Pseudomonas*, and *Salmonella*, these maturases could be activated and the group II introns could be transposed to the *Symbiobacterium* genome. Of course, it is also possible that this transposition of the group II introns is inhibited by another gene silencing system.

It is suggested that *Symbiobacterium* has gained many DNA fragments from phylogenetically distant organisms during the early stage of evolution in the Firmicutes (consisting of Bacilli and Clostridia). As the *Symbiobacterium* genes of all categories have a high GC content (Supplementary Figure 1), it can be concluded that, after acquiring genes, *Symbiobacterium* increased the GC content of the horizontally transferred genes and thereby maintained a genome with a high GC content.

In contrast to the *Symbiobacterium* genome, the *Fusobacterium* (phylogenetically closely related to Firmicutes) genome has a low GC content (27%) [13]. It is suggested that *Fusobacterium* has gained many genes from phylogenetically distant organisms [13]. In the course of evolution, *Fusobacterium* has probably decreased the GC content of the horizontally acquired genes and maintained a genome with a low GC content.

Does *Symbiobacterium* benefit from maintaining a genome with a high GC content? Considering that CO_2 is the major growth factor of *Symbiobacterium*, its symbiotic partners may not be limited to *Geobacillus*. *Symbiobacterium* is widespread in different natural environments [26, 27]. The difference in the genome base compositions between *Symbiobacterium* and its symbiotic partners may lead to a decrease in the frequency of a homologous recombination between the 2 genomes. For example, the 5 sequenced chromosomal genomes of *Geobacillus* have a GC content ranging from 42.8% to 52.5% (http://insilico.ehu.es/oligoweb/).

In addition, homologous recombination is generally effective for adaptive evolution [11]. However, if the population density is low or the recombining population is rare in the environment, adaptive evolution is hampered [11]. Considering the wide distribution of *Symbiobacterium* in natural environments, the population size of *Symbiobacterium* may be adequately large, suggesting that homologous recombination between the *Symbiobacterium* strains and different symbiotic partners may be effective for adaptive evolution. Thus, it is hypothesized that *Symbiobacterium* has maintained its extreme genome composition to avoid homologous recombination between its genome and the genomes of different species and to promote homologous recombination between its genome and the genomes of the same species (or genus).

Acknowledgment

The authors thank Professor Teruhiko Beppu for the helpful comments and critical review of the paper.

References

[1] K. Ueda and T. Beppu, "Lessons from studies of *Symbiobacterium thermophilum*, a unique syntrophic bacterium," *Bioscience, Biotechnology and Biochemistry*, vol. 71, no. 5, pp. 1115–1121, 2007.

[2] H. Nishida, T. Beppu, and K. Ueda, "*Symbiobacterium* lost carbonic anhydrase in the course of evolution," *Journal of Molecular Evolution*, vol. 68, no. 1, pp. 90–96, 2009.

[3] K. Ueda, A. Yamashita, J. Ishikawa et al., "Genome sequence of *Symbiobacterium thermophilum*, an uncultivable bacterium that depends on microbial commensalism," *Nucleic Acids Research*, vol. 32, no. 16, pp. 4937–4944, 2004.

[4] T. O. Watsuji, T. Kato, K. Ueda, and T. Beppu, "CO_2 supply induces the growth of *Symbiobacterium thermophilum*, a syntrophic bacterium," *Bioscience, Biotechnology and Biochemistry*, vol. 70, no. 3, pp. 753–756, 2006.

[5] G. Ding, Z. Yu, J. Zhao et al., "Tree of life based on genome context networks," *PLoS One*, vol. 3, no. 10, Article ID e3357, 2008.

[6] E. P. C. Rocha and E. J. Feil, "Mutational patterns cannot explain genome composition: are there any neutral sites in the genomes of bacteria?" *PLoS Genetics*, vol. 6, no. 9, Article ID e1001104, 2010.

[7] F. Hildebrand, A. Meyer, and A. Eyre-Walker, "Evidence of selection upon genomic GC-content in bacteria," *PLoS Genetics*, vol. 6, no. 9, Article ID e1001107, 2010.

[8] R. Hershberg and D. A. Petrov, "Evidence that mutation is universally biased towards AT in bacteria," *PLoS Genetics*, vol. 6, no. 9, Article ID e1001115, 2010.

[9] J. P. Gogarten and J. P. Townsend, "Horizontal gene transfer, genome innovation and evolution," *Nature Reviews Microbiology*, vol. 3, no. 9, pp. 679–687, 2005.

[10] E. V. Koonin, K. S. Makarova, and L. Aravind, "Horizontal gene transfer in prokaryotes: quantification and classification," *Annual Review of Microbiology*, vol. 55, pp. 709–742, 2001.

[11] B. R. Levin and O. E. Cornejo, "The population and evolutionary dynamics of homologous gene recombination in bacteria," *PLoS Genetics*, vol. 5, no. 8, Article ID e1000601, 2009.

[12] J. M. Smith, N. H. Smith, M. O'Rourke, and B. G. Spratt, "How clonal are bacteria?" *Proceedings of the National Academy of Sciences of the United States of America*, vol. 90, no. 10, pp. 4384–4388, 1993.

[13] A. Mira, R. Pushker, B. A. Legault, D. Moreira, and F. Rodríguez-Valera, "Evolutionary relationships of *Fusobacterium nucleatum* based on phylogenetic analysis and comparative genomics," *BMC Evolutionary Biology*, vol. 4, article no. 50, 2004.

[14] C. A. Fuchsman and G. Rocap, "Whole-genome reciprocal BLAST analysis reveals that *Planctomycetes* do not share an unusually large number of genes with *Eukarya* and *Archaea*," *Applied and Environmental Microbiology*, vol. 72, no. 10, pp. 6841–6844, 2006.

[15] Y. Nakamura, T. Itoh, H. Matsuda, and T. Gojobori, "Biased biological functions of horizontally-transferred genes in prokaryotic genomes," *Nature Genetics*, vol. 36, no. 7, pp. 760–766, 2004.

[16] M. G. I. Langille, W. W. L. Hsiao, and F. S. L. Brinkman, "Detecting genomic islands using bioinformatics approaches," *Nature Reviews Microbiology*, vol. 8, no. 5, pp. 373–382, 2010.

[17] S. Lucchini, G. Rowley, M. D. Goldberg, D. Hurd, M. Harrison, and J. C. Hinton, "H-NS mediates the silencing of laterally acquired genes in bacteria.," *PLoS Pathogens*, vol. 2, no. 8, article e81, 2006.

[18] W. W. Navarre, S. Porwollik, Y. Wang et al., "Selective silencing of foreign DNA with low GC content by the H-NS protein in *Salmonella*," *Science*, vol. 313, no. 5784, pp. 236–238, 2006.

[19] C. -S. Yun, C. Suzuki, K. Naito et al., "Pmr, a histone-like protein H1 (H-NS) family protein encoded by the IncP-7 plasmid pCAR1, is a key global regulator that alters host function," *Journal of Bacteriology*, vol. 192, no. 18, pp. 4720–4731, 2010.

[20] B. R. G. Gordon, Y. Li, L. Wang et al., "Lsr2 is a nucleoid-associated protein that targets AT-rich sequences and virulence genes in *Mycobacterium tuberculosis*," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 107, no. 11, pp. 5154–5159, 2010.

[21] J. G. Lawrence and H. Ochman, "Amelioration of bacterial genomes: rates of change and exchange," *Journal of Molecular Evolution*, vol. 44, no. 4, pp. 383–397, 1997.

[22] N. Sueoka, "On the genetic basis of variation and heterogeneity of DNA base composition," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 48, pp. 582–592, 1962.

[23] T. Oshima, S. Ishikawa, K. Kurokawa, H. Aiba, and N. Ogasawara, "*Escherichia coli* histone-like protein H-NS preferentially binds to horizontally acquired DNA in association with RNA polymerase," *DNA Research*, vol. 13, no. 4, pp. 141–153, 2006.

[24] F. Michel and J. L. Ferat, "Structure and activities of group II introns," *Annual Review of Biochemistry*, vol. 64, pp. 435–461, 1995.

[25] M. Matsuura, J. W. Noah, and A. M. Lambowitz, "Mechanism of maturase-promoted group II intron splicing," *EMBO Journal*, vol. 20, no. 24, pp. 7259–7270, 2002.

[26] T. Sugihara, T. O. Watsuji, S. Kubota et al., "Distribution of *Symbiobacterium thermophilum* and related bacteria in the marine environment," *Bioscience, Biotechnology and Biochemistry*, vol. 72, no. 1, pp. 204–211, 2008.

[27] K. Ueda, M. Ohno, K. Yamamoto et al., "Distribution and diversity of symbiotic thermophiles, *Symbiobacterium thermophilum* and related bacteria, in natural environments," *Applied and Environmental Microbiology*, vol. 67, no. 9, pp. 3779–3784, 2001.

SAGE-Hindawi Access to Research
International Journal of Evolutionary Biology
Volume 2011, Article ID 376831, 8 pages
doi:10.4061/2011/376831

Research Article

Unique Evolution of *Symbiobacterium thermophilum* Suggested from Gene Content and Orthologous Protein Sequence Comparisons

Kenro Oshima,[1] Kenji Ueda,[2] Teruhiko Beppu,[2] and Hiromi Nishida[3]

[1] *Department of Agricultural and Environmental Biology, Graduate School of Agricultural and Life Sciences, The University of Tokyo, Bunkyo-ku, Tokyo 113-8657, Japan*
[2] *Life Science Research Center, College of Bioresource Sciences, Nihon University, 1866 Kameino, Fujisawa 252-8510, Japan*
[3] *Agricultural Bioinformatics Research Unit, Graduate School of Agriculture and Life Sciences, The University of Tokyo, Bunkyo-ku, Tokyo 113-8657, Japan*

Correspondence should be addressed to Hiromi Nishida, hnishida@iu.a.u-tokyo.ac.jp

Received 21 September 2010; Accepted 27 November 2010

Academic Editor: Hideaki Nojiri

Comparisons of gene content and orthologous protein sequence constitute a major strategy in whole-genome comparison studies. It is expected that horizontal gene transfer between phylogenetically distant organisms and lineage-specific gene loss have greater influence on gene content-based phylogenetic analysis than orthologous protein sequence-based phylogenetic analysis. To determine the evolution of the syntrophic bacterium *Symbiobacterium thermophilum*, we analyzed phylogenetic relationships among Clostridia on the basis of gene content and orthologous protein sequence comparisons. These comparisons revealed that these 2 phylogenetic relationships are topologically different. Our results suggest that each Clostridia has a species-specific gene content because frequent genetic exchanges or gene losses have occurred during evolution. Specifically, the phylogenetic positions of syntrophic Clostridia were different between these 2 phylogenetic analyses, suggesting that large diversity in the living environments may cause the observed species-specific gene content. *S. thermophilum* occupied the most distant position from the other syntrophic Clostridia in the gene content-based phylogenetic tree. We identified 32 genes (14 under relaxed selection and 18 under functional constraint) evolving under *Symbiobacterium*-specific selection on the basis of synonymous-to-nonsynonymous substitution ratios. Five of the 14 genes under relaxed selection are related to transcription. In contrast, none of the 18 genes under functional constraint is related to transcription.

1. Introduction

Symbiobacterium thermophilum is a phylogenetically unique bacterium that effectively grows only in coculture with a cognate *Geobacillus* sp. [1]. 16S rDNA-based phylogenetic analysis has shown that it is actually a Gram-positive bacterium [2]. Although *S. thermophilum* phylogenetically belongs to Clostridia (low GC-content bacterial group), the genome of *S. thermophilum* has a high GC content (68.7%) [3]. Furthermore, 2 recent independent analyses concluded that *Symbiobacterium* affiliates with Clostridia (a class of Firmicutes): Ding et al. [4] carried out genome-context network analysis of 195 fully sequenced representative species, including *S. thermophilum*, and we analyzed the concatenated alignment of ribosomal protein sequences [5].

In a previous phylogenetic analysis that was based on ribosomal protein sequence comparisons [5], *S. thermophilum* was closely related to 6 recently sequenced Clostridia that have distinct properties, that is, *Carboxydothermus hydrogenoformans*, *Desulfitobacterium hafniense*, *Moorella thermoacetica*, *Pelotomaculum thermopropionicum*, *Desulfotomaculum reducens*, and *Syntrophomonas wolfei*. *Symbiobacterium* is dependent on the multiple functions of *Geobacillus,* including the supply of CO_2 [1]. *C. hydrogenoformans* [6] grows by utilizing CO as a sole carbon source and water as an electron acceptor, which produces CO_2

and hydrogen as waste products. *D. hafniense* [7] carries out anaerobic dechlorination of tetrachloroethene (PCE). *M. thermoacetica* [8] is an acetogenic bacterium that has been widely used to study the Wood-Ljungdahl pathway of CO and CO_2 fixation (reductive acetyl-CoA pathway). *P. thermopropionicum* [9] is a member of a complex anaerobic microbial consortium where it catalyzes the intermediate bottleneck step by digesting volatile fatty acids (VFAs) and alcohols produced by upstream fermenting bacteria and it supplies acetate, hydrogen, and CO_2 to downstream methanogenic archaea. *D. reducens* is an anaerobic sulfate-reducing bacterium [10]. *S. wolfei* is a fatty-acid-degrading hydrogen/formate-producing anaerobic bacterium [11].

Comparisons of gene content and orthologous protein sequence constitute the major strategy in the whole-genome comparison study [12]. Clostridia have the large amount of bacteria. The phylogenetic position of *Symbiobacterium* remains uncertain in Clostridia. In this study, we reconstructed phylogenetic trees of Clostridia on the basis of the 2 different methods and compared them.

2. Methods

2.1. Phylogenetic Analysis on the Basis of Gene Content Comparisons. We used the following 51 bacteria (50 Clostridia and 1 *Bacillus* belonging to Firmicutes) in this analysis: *Alkaliphilus metalliredigens*, *Alkaliphilus oremlandii*, *Ammonifex degensii*, *Anaerocellum thermophilum*, *Anaerococcus prevotii*, *Bacillus subtilis*, *Caldicellulosiruptor saccharolyticus*, *Candidatus* Desulforudis audaxviator, *Carboxydothermus hydrogenoformans*, *Clostridium acetobutylicum*, *Clostridium beijerinckii*, *Clostridium botulinum* A ATCC 19397, *C. botulinum* A ATCC 3502, *C. botulinum* A Hall, *C. botulinum* A2, *C. botulinum* A3 Loch Maree, *C. botulinum* B Eklund 17B, *C. botulinum* B1 Okra, *C. botulinum* Ba4, *C. botulinum* E3, *C. botulinum* F Langeland, *Clostridium cellulolyticum*, *Clostridium difficile* 630, *C. difficile* CD196, *Clostridium kluyveri* DSM 555, *C. kluyveri* NBRC 12016, *Clostridium novyi*, *Clostridium perfringens* ATCC 13124, *C. perfringens* SM101, *C. perfringens* 13, *Clostridium phytofermentans*, *Clostridium tetani* E88, *Clostridium thermocellum*, *Coprothermobacter proteolyticus*, *Desulfitobacterium hafniense* DCB-2, *D. hafniense* Y51, *Desulfotomaculum acetoxidans*, *Desulfotomaculum reducens*, *Eubacterium eligens*, *Eubacterium rectale*, *Finegoldia magna*, *Halothermothrix orenii*, *Heliobacterium modesticaldum*, *Moorella thermoacetica*, *Natranaerobius thermophilus*, *Pelotomaculum thermopropionicum*, *Symbiobacterium thermophilum*, *Syntrophomonas wolfei*, *Thermoanaerobacter pseudethanolicus*, *Thermoanaerobacter* sp. X514, and *Thermoanaerobacter tengcongensis*. Ortholog cluster analysis among the above 51 bacteria was performed using the MBGD [13] (Microbial Genome Database for Comparative Analysis; http://mbgd.nibb.ac.jp/). The analysis (minimum cluster size, 2) provided a gene presence/absence data matrix (10,636 genes × 51 organisms), which served as the basis for a distance matrix between all pairs of the 51 organisms. The distance was calculated from the different ratios between the presence/absence patterns of the 10,636 genes. On the basis of distance matrix, a neighbor-joining tree was reconstructed using MEGA software version 4 [14]. The bootstrap was performed with 1000 replicates.

2.2. Phylogenetic Analysis on the Basis of 112 Orthologous Protein Sequence Comparisons. We used the following 55 bacteria (54 Clostridia and 1 *Bacillus*) in this analysis: *Acidaminococcus fermentans*, *A. metalliredigens*, *A. degensii*, *A. thermophilum*, *A. prevotii*, *B. subtilis*, *C. saccharolyticus*, *Candidatus* D. audaxviator, *C. hydrogenoformans*, Clostridiales genomosp. BVAB3 UPII9-5, *C. acetobutylicum*, *C. beijerinckii*, *C. botulinum* A ATCC 19397, *C. botulinum* A ATCC 3502, *C. botulinum* A Hall, *C. botulinum* A2 Kyoto, *C. botulinum* A3 Loch Maree, *C. botulinum* B Eklund 17B, *C. botulinum* B1 Okra, *C. botulinum* Ba4 657, *C. botulinum* E3 Alaska E43, *C. botulinum* F Langeland, *C. cellulolyticum*, *C. difficile* 630, *C. difficile* CD196, *C. difficile* R20291, *C. kluyveri* DSM 555, *C. kluyveri* NBRC 12016, *C. novyi*, *C. perfringens* ATCC 13124, *C. perfringens* SM101, *C. perfringens* 13, *C. phytofermentans*, *C. tetani*, *C. thermocellum*, *C. proteolyticus*, *D. hafniense* DCB-2, *D. hafniense* Y51, *D. acetoxidans*, *D. reducens*, *E. eligens*, *E. rectale*, *F. magna*, *H. orenii*, *H. modesticaldum*, *M. thermoacetica*, *N. thermophilus*, *P. thermopropionicum*, *S. thermophilum*, *S. wolfei*, *Thermoanaerobacter italicus*, *T. pseudethanolicus*, *T.* sp. X514, *T. tengcongensis*, and *Veillonella parvula*. From the above 55 bacteria, 112 proteins were extracted as orthologous proteins by using a previously described method [15]. Thus, we constructed 112 multiple alignments using Clustal W [16]. Then, a concatenated multiple alignment of the 112 multiple alignments was generated. The complete multiple alignment had 52,204 amino acid sites, including 19,818 gap/insertion sites. Hence, phylogenetic analyses were performed on the basis of 32,386 amino acid sites without the gap/insertion sites. The neighbor-joining tree was reconstructed using MEGA software version 4 [14]. The bootstrap was performed with 1000 replicates. The rate variation among sites was considered to have a gamma-distributed rate ($\alpha = 1$). The other default parameters (e.g., Poisson distance) were not changed.

2.3. Extraction of Genes Evolving under Symbiobacterium–Specific Selection among Syntrophic Clostridia. Among *Bacillus subtilis*, *Carboxydothermus hydrogenoformans*, *Desulfitobacterium hafniense*, *Moorella thermoacetica*, *Pelotomaculum thermopropionicum*, *Desulfotomaculum reducens*, *Symbiobacterium thermophilum*, and *Syntrophomonas wolfei*, 472 genes were extracted as orthologous genes by the previously described method [15]. Synonymous substitution occurs more frequently than nonsynonymous substitution in protein-coding sequences because of relaxed functional constraints (nonsynonymous-to-synonymous ratio $\omega < 1$) [17], whereas they occur equally in noncoding regions and pseudogenes ($\omega = 1$). We calculated the likelihood of both the codon substitution model allowing for one ω (model R1) and the *S. thermophilum* branch-specific model allowing for 2 ratios (ω_0 and ω_1; model R2), using PAML version 3.14 [18]. In model R2, the branches of the gene tree were partitioned into the *Symbiobacterium* branch (ω_1) and other related branches (ω_0). Likelihood ratio test statistics were calculated as twice the difference between the 2 log-likelihoods (2Δ ln) and compared with a χ^2 distribution with degrees of freedom equal to the difference in the number of parameters between the 2 models [19]. According to this method, the genes evolving under the *Symbiobacterium*-specific selection among *Bacillus* and 7 Clostridia were extracted.

3. Results and Discussion

Phylogenetic relationships among Clostridia on the basis of gene content comparison (Figure 1) were topologically different from those generated on the basis of orthologous protein sequence comparison (Figure 2). For example, in the gene content-based phylogenetic tree, *Alkaliphilus*, *Clostridium* (except for *C. cellulolyticum* and *C. thermocellum*), *Desulfitobacterium*, and *Eubacterium* formed a monophyletic lineage with 85% bootstrap support (Figure 1). In contrast, in the 112 orthologous protein sequence-based phylogenetic relationships, *Alkaliphilus*, *Anaerococcus*, *Clostridium* (except for *C. cellulolyticum* and *C. thermocellum*), *Eubacterium*, and *Finegoldia* formed a monophyletic lineage with 98% bootstrap support (Figure 2). Thus, the phylogenetic positions of *Anaerococcus*, *Desulfitobacterium*, and *Finegoldia* were different between these 2 trees. In addition, *Coprothermobacter proteolyticus* was positioned differently in the 2 trees. Moreover, the very long branch in the orthologous protein-based tree suggests that *C. proteolyticus* has a substitution pattern that is different from other related Clostridia.

We expected horizontal gene transfer between phylogenetically distant organisms and lineage-specific gene loss to have greater influence on the gene content-based phylogenetic analysis than the orthologous protein-based analysis [12, 20]. Bacteria make their gene content suitable for the living environment by changing it through gene acquisition and loss.

The phylogenetic positions of 2 *D. hafniense* strains are located near those of *Alkaliphilus*, *Clostridium* (except for *C. cellulolyticum* and *C. thermocellum*), and *Eubacterium* in the gene content-based phylogenetic tree (Figure 1). However, those phylogenetic positions were located in the phylogenetic lineage of syntrophic Clostridia in the orthologous protein-based tree (Figure 2). The gene content-based phylogenetic tree (Figure 1) indicates that *Symbiobacterium* branched off at the earliest stage of Clostridia species diversification. In contrast, *Natranaerobius* branched off at the earliest species diversification stage in the orthologous protein sequence-based phylogenetic tree (Figure 2).

Although *S. thermophilum* occupied the most basal position in the gene content-based Clostridia lineage (Figure 1), it was located in the syntrophic Clostridia lineage on the basis of orthologous protein sequence comparisons (Figure 2). Syntrophic bacteria evolved to acquire different sets of genes despite their close phylogenetic relationship. Thus, although *Symbiobacterium* clusters with syntrophic Clostridia, its gene content is very different. *S. thermophilum* has the most distant position from the other syntrophic Clostridia in the phylogenetic tree on the basis of gene content comparisons.

Although the physiological reason for the high CO_2 requirement of *S. thermophilum* is not yet known, we assumed that it is related to the carbonic anhydrase deficiency (the ubiquitous enzyme catalyzing interconversion between CO_2 and bicarbonate; EC 4.2.1.1), as deficiency of this enzyme results in the need for high CO_2 levels in several model microorganisms [1]. *S. thermophilum* lost this enzyme in the course of evolution [5]. In this previous analysis, we inferred that *C. hydrogenoformans* and *M. thermoacetica* have also lost the gene for carbonic anhydrase; however, we recently noticed that *C. hydrogenoformans* had 2 potential carbonic anhydrase coding genes with structures different from the other syntrophic Clostridia carbonic anhydrases. Therefore, only *Moorella* has lost the carbonic anhydrase gene, in addition to *Symbiobacterium*. However, according to our results, these two bacteria are not closely related to each other (Figures 1 and 2), suggesting that the gene loss in these 2 species occurred independently during evolution.

Our results imply that each syntrophic Clostridial organism, especially *Symbiobacterium*, would have genes that evolved in an organism-specific manner. We expect that characterization of such genes will provide useful information with regard to the evolutionary history and physiological features specific to the corresponding organism [21, 22]. We identified 32 genes evolving under *Symbiobacterium*-specific selection (Table 1). The analysis revealed that the likelihood of model R2 was significantly higher ($P < .05$) than that of model R1 in the 32 genes. Of these, 14 genes showed $\omega_1/\omega_0 > 1$ and 18 showed $\omega_1/\omega_0 < 1$.

Among the 32 genes evolving under *Symbiobacterium*-specific selection, the RNA chaperone Hfq-coding gene has the highest ω_1 value (0.5347) (Table 1). Hfq facilitates pairing interactions between small regulatory RNAs and their mRNA targets, which has a variety of functions in bacteria [23]. Among 73 conserved amino acid sites of Hfq (Figure 3), *S. thermophilum* has more specific sites (7 sites) than the outgroup *Bacillus* (4 sites), indicating that the Hfq gene is one of the genes evolving under *Symbiobacterium*-specific selection.

Two genes related to transcription, *sigA* (RNA polymerase sigma factor coding gene) and *rpoC* (RNA polymerase subunit beta' coding gene) have evolved under relaxed selection (Table 1). These results could be related to the high GC content of *Symbiobacterium* genes. Thus, we hypothesized that the GC bias of the promoter sequence induced *Symbiobacterium*-specific SigA, a DNA-binding protein, which led to the structural change of RNA polymerase complex (including RpoC). We discussed the relationships between the GC content and phylogeny of the *Symbiobacterium* genes [24].

In addition, *spoIIAB* and *cheY* are also related to transcription. Thus, 5 of the 14 genes under more relaxed selection than other Clostridia are related to transcription. However, none of the 18 genes under functional constraint is related to transcription. Those results suggest that, under relaxed selection, the transcription system may be related to *S. thermophilum*-specific gene content. In fact, *Symbiobacterium* lost the transcriptional regulator genes *arsR*, *GntR*, and *Lrp* compared to other syntrophic Clostridia

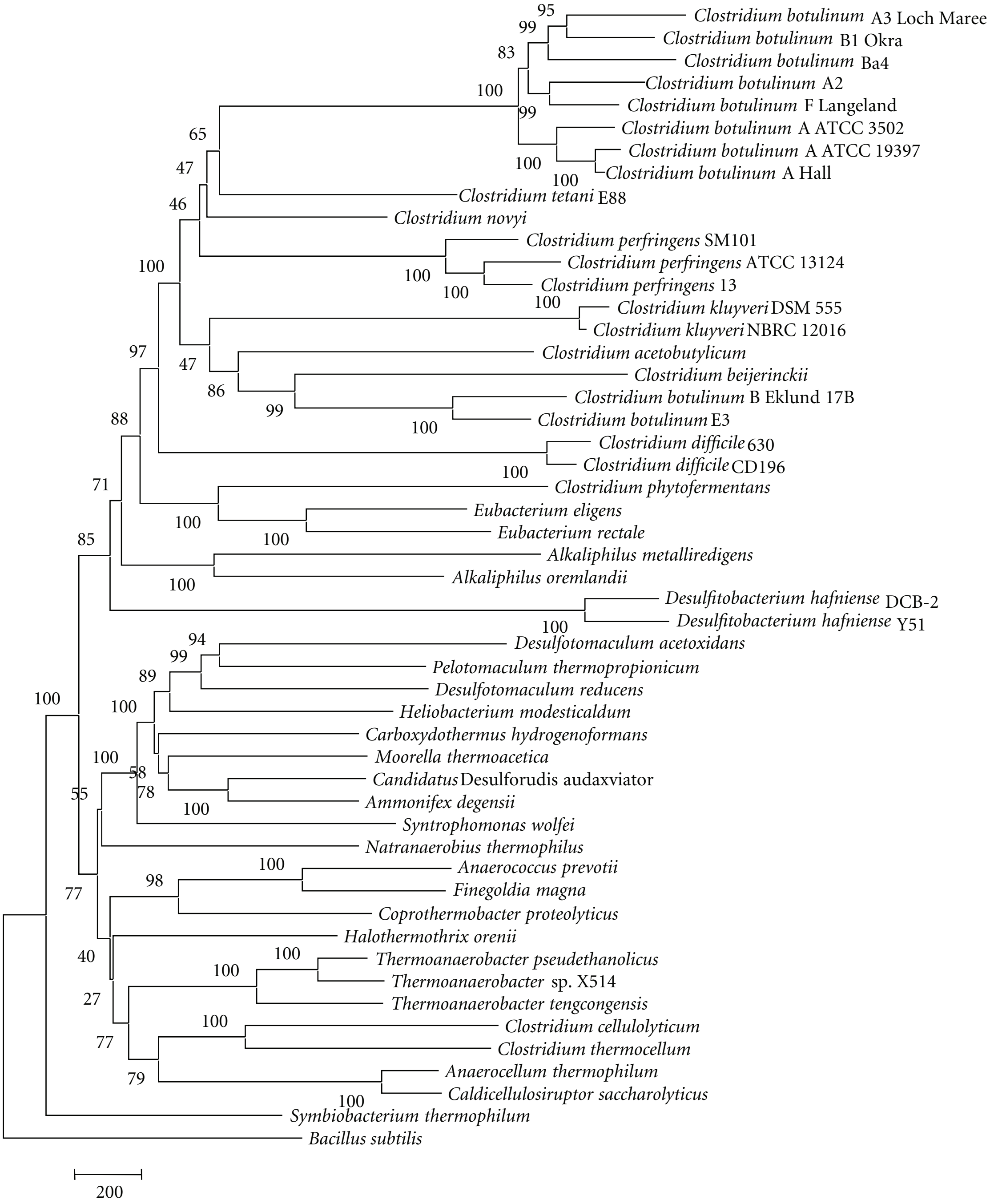

95
99
83
100
99
100
100
65
47
46
100
100
100
100
97
47
86
99
100
100
88
71
100
100
85
100
100
94
99
89
100
100
58
78
100
55
100
98
77
40
100
100
27
100
77
79
100
Clostridium botulinum A3 Loch Maree
Clostridium botulinum B1 Okra
Clostridium botulinum Ba4
Clostridium botulinum A2
Clostridium botulinum F Langeland
Clostridium botulinum A ATCC 3502
Clostridium botulinum A ATCC 19397
Clostridium botulinum A Hall
Clostridium tetani E88
Clostridium novyi
Clostridium perfringens SM101
Clostridium perfringens ATCC 13124
Clostridium perfringens 13
Clostridium kluyveri DSM 555
Clostridium kluyveri NBRC 12016
Clostridium acetobutylicum
Clostridium beijerinckii
Clostridium botulinum B Eklund 17B
Clostridium botulinum E3
Clostridium difficile 630
Clostridium difficile CD196
Clostridium phytofermentans
Eubacterium eligens
Eubacterium rectale
Alkaliphilus metalliredigens
Alkaliphilus oremlandii
Desulfitobacterium hafniense DCB-2
Desulfitobacterium hafniense Y51
Desulfotomaculum acetoxidans
Pelotomaculum thermopropionicum
Desulfotomaculum reducens
Heliobacterium modesticaldum
Carboxydothermus hydrogenoformans
Moorella thermoacetica
Candidatus Desulforudis audaxviator
Ammonifex degensii
Syntrophomonas wolfei
Natranaerobius thermophilus
Anaerococcus prevotii
Finegoldia magna
Coprothermobacter proteolyticus
Halothermothrix orenii
Thermoanaerobacter pseudethanolicus
Thermoanaerobacter sp. X514
Thermoanaerobacter tengcongensis
Clostridium cellulolyticum
Clostridium thermocellum
Anaerocellum thermophilum
Caldicellulosiruptor saccharolyticus
Symbiobacterium thermophilum
Bacillus subtilis
200

Figure 1: Phylogenetic relationships on the basis of gene content comparisons among 50 Clostridia and *Bacillus subtilis*. The ortholog cluster analysis (minimum cluster size, 2) among the 51 bacteria was performed using the MBGD [13]. This analysis produced the gene presence/absence data matrix (10,636 genes × 51 organisms), which was used to generate the distance matrix between all pairs of the 51 bacteria. On the basis of the distance matrix, a neighbor-joining tree was reconstructed using MEGA software version 4 [14]. The bootstrap was performed with 1000 replicates. The bar indicates a 200-gene difference.

(See in the Supplementary Material available online at doi: 10.4061/2011/376831 Table S1.).

It is noteworthy that some functionally related genes exhibited opposite nucleotide substitution patterns in *S. thermophilum* (Table 1). For example, *argD* (*N*-acetylornithine aminotransferase coding gene) has evolved under relaxed selection whereas *argC* (*N*-acetyl-gamma-glutamyl-phosphate reductase coding gene) has evolved under functional constraint. Another example is the genes encoding flagella-associated proteins; *flgG* (flagellar hook protein coding gene) has evolved under relaxed selection, whereas *flgD* (flagellar hook assembly protein coding gene)

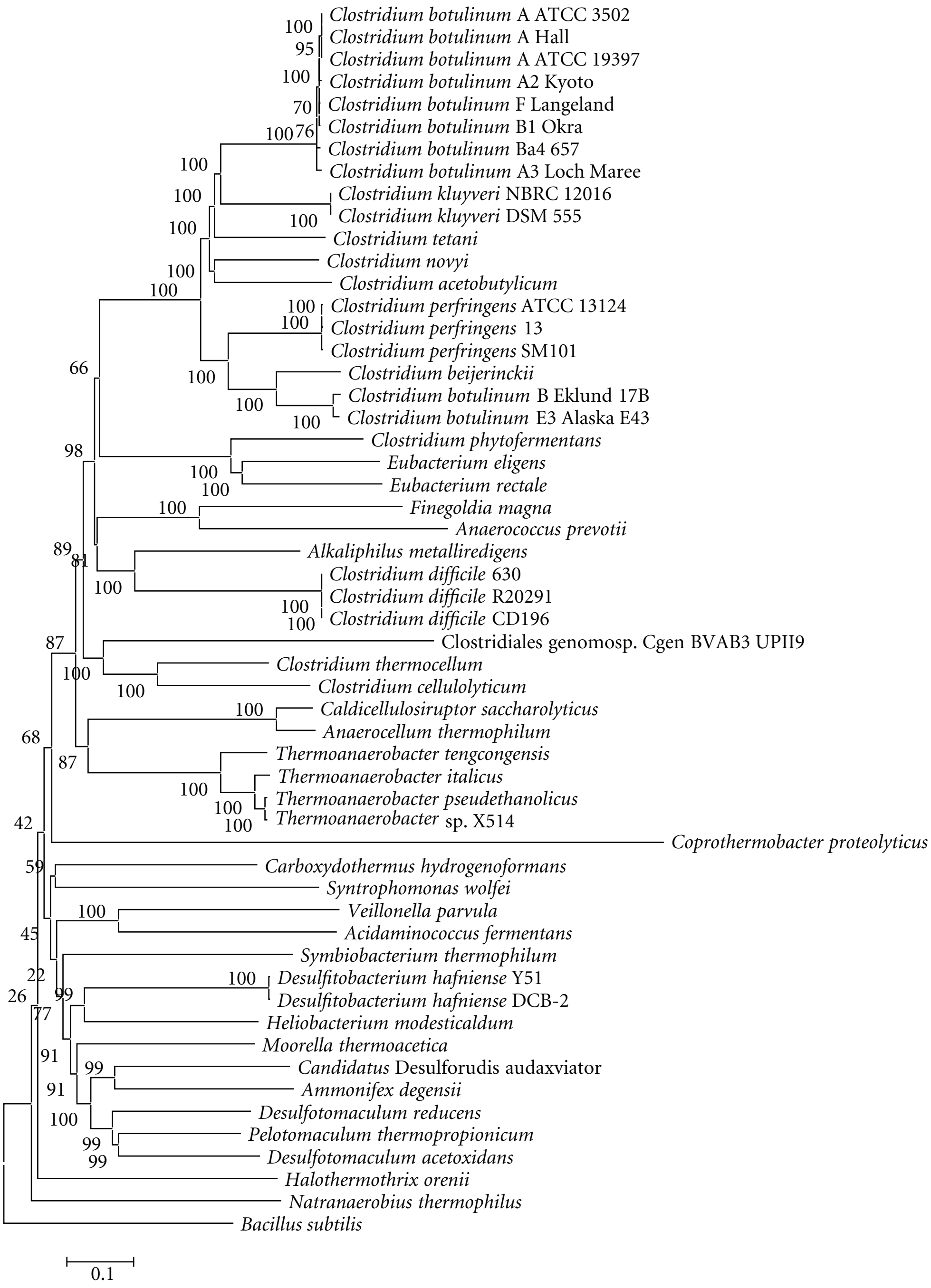

Clostridium botulinum A ATCC 3502
Clostridium botulinum A Hall
Clostridium botulinum A ATCC 19397
Clostridium botulinum A2 Kyoto
Clostridium botulinum F Langeland
Clostridium botulinum B1 Okra
Clostridium botulinum Ba4 657
Clostridium botulinum A3 Loch Maree
Clostridium kluyveri NBRC 12016
Clostridium kluyveri DSM 555
Clostridium tetani
Clostridium novyi
Clostridium acetobutylicum
Clostridium perfringens ATCC 13124
Clostridium perfringens 13
Clostridium perfringens SM101
Clostridium beijerinckii
Clostridium botulinum B Eklund 17B
Clostridium botulinum E3 Alaska E43
Clostridium phytofermentans
Eubacterium eligens
Eubacterium rectale
Finegoldia magna
Anaerococcus prevotii
Alkaliphilus metalliredigens
Clostridium difficile 630
Clostridium difficile R20291
Clostridium difficile CD196
Clostridiales genomosp. Cgen BVAB3 UPII9
Clostridium thermocellum
Clostridium cellulolyticum
Caldicellulosiruptor saccharolyticus
Anaerocellum thermophilum
Thermoanaerobacter tengcongensis
Thermoanaerobacter italicus
Thermoanaerobacter pseudethanolicus
Thermoanaerobacter sp. X514
Coprothermobacter proteolyticus
Carboxydothermus hydrogenoformans
Syntrophomonas wolfei
Veillonella parvula
Acidaminococcus fermentans
Symbiobacterium thermophilum
Desulfitobacterium hafniense Y51
Desulfitobacterium hafniense DCB-2
Heliobacterium modesticaldum
Moorella thermoacetica
Candidatus Desulforudis audaxviator
Ammonifex degensii
Desulfotomaculum reducens
Pelotomaculum thermopropionicum
Desulfotomaculum acetoxidans
Halothermothrix orenii
Natranaerobius thermophilus
Bacillus subtilis
100
95
100
70
100
76
100
100
100
100
100
100
100
100
100
100
100
66
100
98
100
100
100
89
81
100
100
100
87
100
100
100
68
87
100
100
100
42
59
45
100
22
26
100
99
77
91
99
91
100
99
99
0.1

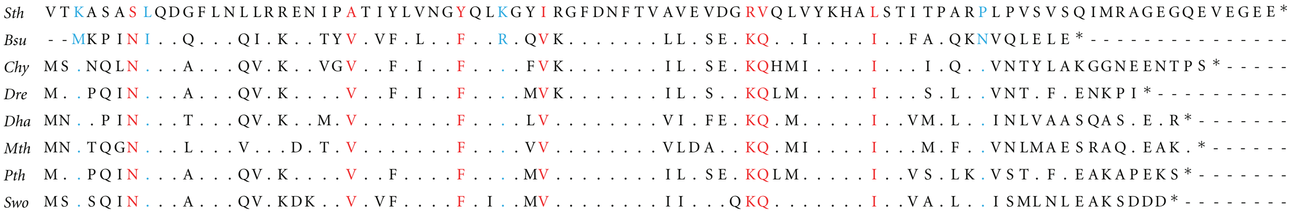
Figure 2: Phylogenetic relationships on the basis of 112 orthologous protein sequence comparisons among 54 Clostridia and *B. subtilis*. The 112 proteins were extracted as orthologous proteins from the 55 bacteria by a previously described method [15]. We constructed the 112 multiple alignments by using Clustal W [16]. Then, a concatenated multiple alignment of the 112 multiple alignments was generated. The complete multiple alignment had 52,204 amino acid sites, including 19,818 gap/insertion sites. Hence, phylogenetic analyses were performed on the basis of 32,386 amino acid sites without the gap/insertion sites. The neighbor-joining tree was reconstructed using MEGA software version 4 [14]. The bootstrap was performed with 1000 replicates. The rate variation among sites was assumed to have a gamma distributed rate ($\alpha = 1$). No other default parameters were changed. The bar indicates a 10% difference.

and *fliS* (flagellar protein coding gene) have evolved under functional constraint. *flgG* exhibited the highest ω_1/ω_0 value (75.48) (Table 1). Flagella mediate interactions between *P. thermopropionicum* and methanogenic archaea [25]. Similar specialized functions in syntrophic association could have been a limiting factor for the evolution of the above 2 flagellum genes in *Symbiobacterium.*

In conclusion, our results suggest that *S. thermophilum* has evolved in a unique manner compared to other syntrophic Clostridia from the perspective of gene content.

Table 1: Genes evolving under *Symbiobacterium*-specific selection.

Gene	ω_1	ω_1/ω_0	$2\Delta\ln$
$\omega_1/\omega_0 > 1$			
hfq (RNA chaperone, STH1746)	0.5347	24.3046	5.7413
spoIIAB (anti-sigma F factor, STH1813)	0.3967	5.9744	8.7835
flgG (flagellar hook protein, STH2995)	0.3774	75.4800	6.4323
ilvC (ketol-acid reductoisomerase, STH2688)	0.2240	3.4675	10.7272
rplL (50S ribosomal protein L7/L12, STH3086)	0.2183	8.3640	13.4750
argD (*N*-acetylornithine aminotransferase, STH2881)	0.2084	2.0292	4.1224
rplK (50S ribosomal protein L11, STH3090)	0.1869	9.3450	4.2192
ylmE (alanine racemase domain-containing protein, STH1227)	0.1526	24.2222	15.4681
proJ (gamma-glutamyl kinase, STH2540)	0.1497	26.2632	4.4715
sigA (RNA polymerase sigma factor, STH0588)	0.1315	3.7679	17.9996
rpoC (RNA polymerase subunit beta', STH3084)	0.0838	2.1487	7.2876
glmS (glucosamine-fructose-6-phosphate aminotransferase, STH1279)	0.0156	2.7857	13.0700
aroE (3-phosphoshikimate 1 carboxyvinyltransferase, STH1419)	0.0125	2.8409	4.4748
cheY (two-component response regulator involved in modulation of flagellar, STH1540)	0.0044	2.9333	6.7786
$\omega_1/\omega_0 < 1$			
flgD (flagellar hook assembly protein, STH2996)	0.0123	0.0715	4.4609
fliS (flagellar protein FliS, STH2976)	0.0073	0.0885	4.0842
yloM (ribosomal RNA small subunit methyltransferase B, STH1349)	0.0045	0.0441	12.0081
ftsH (cell division protease, STH3198)	0.0040	0.0655	11.9908
spoVFB (dipicolinate synthase subunit B, STH1546)	0.0039	0.0591	6.5852
rplW (50S ribosomal protein L23, STH3073)	0.0039	0.1429	3.9835
trmD (tRNA methyltransferase, STH1470)	0.0038	0.0574	5.7865
argC (*N*-acetyl-gamma-glutamyl-phosphate reductase, STH2892)	0.0038	0.0721	4.1368
rpsC (30S ribosomal protein S3, STH3069)	0.0037	0.1504	4.1064
prfA (peptide chain release factor RF-1, STH0073)	0.0035	0.0750	6.3618
ligA (NAD-dependent DNA ligase, STH2825)	0.0034	0.0654	4.5717
spo0J (ParB-like nuclease domain-containing protein, STH3332)	0.0034	0.0397	10.1363
ftsE (cell division ATP-binding protein, STH0139)	0.0027	0.0407	5.6285
metG (methionyl-tRNA synthetase, STH3252)	0.0027	0.0470	4.1885
rplC (50S ribosomal protein L3, STH3075)	0.0023	0.0920	4.3304
rplB (50S ribosomal protein L2, STH3072)	0.0020	0.0617	3.9836
rpsH (30S ribosomal protein S8, STH3061)	0.0018	0.1047	4.5829
infA (translation initiation factor IF-1, STH3052)	0.0004	0.0234	4.7842



```
Sth  V T K A S A S L Q D G F L N L L R R E N I P A T I Y L V N G Y Q L K G Y I R G F D N F T V A V E V D G R V Q L V Y K H A L S T I T P A R P L P V S V S Q I M R A G E G Q E V E G E E *
Bsu  - - M K P I N I . . Q . . . Q I . K . . T Y V . V F . L . . F . . R . Q V K . . . . . . . L L . S E . K Q . . I . . . . I . . F A . Q K N V Q L E L E * - - - - - - - - - - - - - - -
Chy  M S . N Q L N . . . A . . . Q V . K . . V G V . . F . I . . F . . . . F V K . . . . . . . I L . S E . K Q H M I . . . . I . . . I . Q . . V N T Y L A K G G N E E N T P S * - - - - -
Dre  M . . P Q I N . . . A . . . Q V . K . . . . V . . F . I . . F . . . . M V K . . . . . . . I L . S . . K Q L M . . . . . I . . . S . L . . V N T . F . E N K P I * - - - - - - - - - -
Dha  M N . . P I N . . . T . . . Q V . K . . M . V . . . . . . . F . . . . L V . . . . . . . . V I . F E . K Q . M . . . . . I . . V M . L . . I N L V A A S Q A S . E . R * - - - - - - -
Mth  M N . T Q G N . . . L . . . V . . . D . T . V . . . . . . . F . . . . V V . . . . . . . . V L D A . . K Q . M I . . . . I . . . M . F . . V N L M A E S R A Q . E A K . * - - - - - -
Pth  M . . P Q I N . . . A . . . Q V . K . . . . V . . F . . . . F . . . . M V . . . . . . . . I L . S E . K Q L M . . . . . I . . V S . L K . V S T . F . E A K A P E K S * - - - - - - -
Swo  M S . S Q I N . . . A . . . Q V . K D K . . V . V F . . . . F . I . . M V . . . . . . . . I I . . . Q K Q . . . . . . . I . . V A . L . . I S M L N L E A K S D D D * - - - - - - - -
```

Figure 3: Alignment of amino acid sequences of Hfq. *Sth, Symbiobacterium thermophilum*; *Bsu, Bacillus subtilis*; *Chy, Carboxydothermus hydrogenoformans*; *Dre, Desulfotomaculum reducens*; *Dha, Desulfitobacterium hafniense*; *Mth, Moorella thermoacetica*; *Pth, Pelotomaculum thermopropionicum*; *Swo, Syntrophomonas wolfei*. Red and blue sites indicate *Symbiobacterium*- and *Bacillus*-specific sites, respectively. The dots represent identical residues of *S. thermophilum* amino acid.

Codon substitution analysis also suggests several unique genes that evolved in a *Symbiobacterium*-specific manner. Although speculative, the gene loss or relaxed evolution of several transcriptional regulator genes implies that environmental response might be involved in *Symbiobacterium*-specific evolution.

Acknowledgment

This study was supported by the High-Tech Research Center Project of the Ministry of Education, Culture, Sports, Science and Technology, Japan.

References

[1] K. Ueda and T. Beppu, “Lessons from studies of *Symbiobacterium thermophilum*, a unique syntrophic bacterium,” *Bioscience, Biotechnology and Biochemistry*, vol. 71, no. 5, pp. 1115–1121, 2007.

[2] M. Ohno, H. Shiratori, M. J. Park et al., “*Symbiobacterium thermophilum* gen. nov., sp. nov., a symbiotic thermophile that depends on co-culture with a *Bacillus* strain for growth,” *International Journal of Systematic and Evolutionary Microbiology*, vol. 50, no. 5, pp. 1829–1832, 2000.

[3] K. Ueda, A. Yamashita, J. Ishikawa et al., “Genome sequence of *Symbiobacterium thermophilum*, an uncultivable bacterium that depends on microbial commensalism,” *Nucleic Acids Research*, vol. 32, no. 16, pp. 4937–4944, 2004.

[4] G. Ding, Z. Yu, J. Zhao et al., “Tree of life based on genome context networks,” *PLoS One*, vol. 3, no. 10, Article ID e3357, 2008.

[5] H. Nishida, T. Beppu, and K. Ueda, “*Symbiobacterium* lost carbonic anhydrase in the course of evolution,” *Journal of Molecular Evolution*, vol. 68, no. 1, pp. 90–96, 2009.

[6] M. Wu, Q. Ren, A. S. Durkin et al., “Life in hot carbon monoxide: the complete genome sequence of *Carboxydothermus hydrogenoformans* Z-2901,” *PLoS Genetics*, vol. 1, no. 5, p. e65, 2005.

[7] H. Nonaka, G. Keresztes, Y. Shinoda et al., “Complete genome sequence of the dehalorespiring bacterium *Desulfitobacterium hafniense* Y51 and comparison with *Dehalococcoides ethenogenes* 195,” *Journal of Bacteriology*, vol. 188, no. 6, pp. 2262–2274, 2006.

[8] E. Pierce, G. Xie, R. D. Barabote et al., “The complete genome sequence of *Moorella thermoacetica* (f. *Clostridium thermoaceticum*),” *Environmental Microbiology*, vol. 10, no. 10, pp. 2550–2573, 2008.

[9] T. Kosaka, S. Kato, T. Shimoyama, S. Ishii, T. Abe, and K. Watanabe, “The genome of *Pelotomaculum thermopropionicum* reveals niche-associated evolution in anaerobic microbiota,” *Genome Research*, vol. 18, no. 3, pp. 442–448, 2008.

[10] B. M. Tebo and A. Y. Obraztsova, “Sulfate-reducing bacterium grows with Cr(VI), U(VI), Mn(IV), and Fe(III) as electron acceptors,” *FEMS Microbiology Letters*, vol. 162, no. 1, pp. 193–198, 1998.

[11] M. J. McInerney, M. P. Bryant, R. B. Hespell, and J. W. Costerton, “*Syntrophomonas wolfei* gen. nov. sp. nov., an anaerobic, syntrophic, fatty acid-oxidizing bacterium,” *Applied and Environmental Microbiology*, vol. 41, no. 4, pp. 1029–1039, 1981.

[12] Y. I. Wolf, I. B. Rogozin, N. V. Grishin, and E. V. Koonin, “Genome trees and the tree of life,” *Trends in Genetics*, vol. 18, no. 9, pp. 472–479, 2002.

[13] I. Uchiyama, T. Higuchi, and M. Kawai, “MBGD update 2010: toward a comprehensive resource for exploring microbial genome diversity,” *Nucleic Acids Research*, vol. 38, no. 1, pp. D361–D365, 2010.

[14] K. Tamura, J. Dudley, M. Nei, and S. Kumar, “MEGA4: molecular evolutionary genetics analysis (MEGA) software version 4.0,” *Molecular Biology and Evolution*, vol. 24, no. 8, pp. 1596–1599, 2007.

[15] K. Oshima and H. Nishida, “Phylogenetic relationships among mycoplasmas based on the whole genomic information,” *Journal of Molecular Evolution*, vol. 65, no. 3, pp. 249–258, 2007.

[16] J. D. Thompson, D. G. Higgins, and T. J. Gibson, “CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice,” *Nucleic Acids Research*, vol. 22, no. 22, pp. 4673–4680, 1994.

[17] W. H. Li, T. Gojobori, and M. Nei, “Pseudogenes as a paradigm of neutral evolution,” *Nature*, vol. 292, no. 5820, pp. 237–239, 1981.

[18] Z. Yang, “PAML: a program package for phylogenetic analysis by maximum likelihood,” *Computer Applications in the Biosciences*, vol. 13, no. 5, pp. 555–556, 1997.

[19] N. Goldman and Z. Yang, “A codon-based model of nucleotide substitution for protein-coding DNA sequences,” *Molecular Biology and Evolution*, vol. 11, no. 5, pp. 725–736, 1994.

[20] B. Snel, M. A. Huynen, and B. E. Dutilh, “Genome trees and the nature of genome evolution,” *Annual Review of Microbiology*, vol. 59, pp. 191–209, 2005.

[21] K. Oshima and H. Nishida, “Detection of the genes evolving under *Ureaplasma*-specific selection,” *Journal of Molecular Evolution*, vol. 66, no. 5, pp. 529–532, 2008.

[22] H. Nishida, "*Ureaplasma* urease genes have undergone a unique evolutionary process," *Open Systems Biology Journal*, vol. 2, pp. 1–7, 2009.

[23] A. Jousselin, L. Metzinger, and B. Felden, "On the facultative requirement of the bacterial RNA chaperone, Hfq," *Trends in Microbiology*, vol. 17, no. 9, pp. 399–405, 2009.

[24] H. Nishida and C.-S. Yun, "Phylogenetic and guanine-cytosine content analysis of *Symbiobacterium thermophilum* genes," *International Journal of Evolutionary Biology*, vol. 2011, p. 5, 2011.

[25] T. Shimoyama, S. Kato, S. Ishii, and K. Watanabe, "Flagellum mediates symbiosis," *Science*, vol. 323, no. 5921, p. 1574, 2009.

SAGE-Hindawi Access to Research
International Journal of Evolutionary Biology
Volume 2011, Article ID 937434, 6 pages
doi:10.4061/2011/937434

Review Article

Prevalence of *Mycobacterium tuberculosis* in Taiwan: A Model for Strain Evolution Linked to Population Migration

Horng-Yunn Dou,[1] Shu-Chen Huang,[1] and Ih-Jen Su[1, 2]

[1] *Division of Infectious Diseases, National Health Research Institutes, No. 35, Keyan Road, Zhunan Town, Miaoli County 350, Taiwan*
[2] *Department of Pathology, National Cheng Kung University Hospital, Tainan 704, Taiwan*

Correspondence should be addressed to Ih-Jen Su, suihjen@nhri.org.tw

Received 14 October 2010; Accepted 20 December 2010

Academic Editor: Shinji Kondo

The global evolution and spread of *Mycobacterium tuberculosis* (MTB), one of the most successful bacterial pathogens, remain a mystery. Advances in molecular technology in the past decade now make it possible to understand MTB strain evolution and transmission in the context of human population migration. Taiwan is a relatively isolated island, serving as a mixing vessel over the past four centuries as colonization by different waves of ethnic groups occurred. By using mycobacterial tandem repeat sequences as genetic markers, the prevalence of MTB strains in Taiwan revealed an interesting association with historical migrations of different ethnic populations, thus providing a good model to explore the global evolution and spread of MTB.

1. Introduction

Tuberculosis (TB) remains a major worldwide health concern and has been characterized as one of three epidemics by the World Health Organization [1]. In 2006, more than 1.5 million people died of TB, an estimated 9.1 million new cases appeared, and the number of total TB cases worldwide reached about 14 million [2]. Findings from sites representative of Neolithic Europe, ancient Egypt, and the Greek and Roman empires revealed that TB is an ancient human disease [3]. Population migration due to wars and New World expedition accounts for the major transmission patterns of microbial pathogens, including *Mycobacterium tuberculosis* (MTB). In the past decade, the prevalence of MTB strains in different geographic regions and ethnic populations has been explored by molecular methods [4–6]. The reports revealed interesting patterns of strain distribution in different ethnic populations, which matched well to historical population migrations [5, 6]. Therefore, strain variations in different populations may be used to elucidate the transmission patterns of MTB.

The distribution of TB in different geographic regions is characterized by the prevalence of different MTB strains with varied virulence and drug resistance. Both environmental and host factors are responsible for the transmission and prevalence of different MTB strains. Because MTB has no detectable horizontal gene transfer [7, 8], large sequence polymorphisms (LPSs) can be used as phylogenetic markers to trace the evolutionary relationships of different strain families. Hirsh et al. presented a phylogenetic analysis of genomic deletions or LSPs, which were identified by comparative genome hybridization using DNA microarrays [7]. Mycobacterial interspersed repetitive units (MIRUs) loci comprise variable numbers of tandem repeat (VNTR) sequences, which allow them to be used as powerful genotyping markers [9]. In terms of genetic diversity and mutation rates, they resemble human microsatellites, which are widely used in human population genetics studies. By conducting MIRU-VNTR typing, Supply et al. were able to detect strong linkage disequilibrium between allele variants at these loci, indicative of a predominant clonal evolution in the MTB complex [8].

Taiwan is a relatively isolated island situated to the southeast of mainland China. The ethnic populations of Taiwan include Han Chinese who migrated to the island in two major waves: the first during the Ming dynasty around 1600 and the second between 1945 and 1950, when members of the military, veterans, and some civilians emigrated from

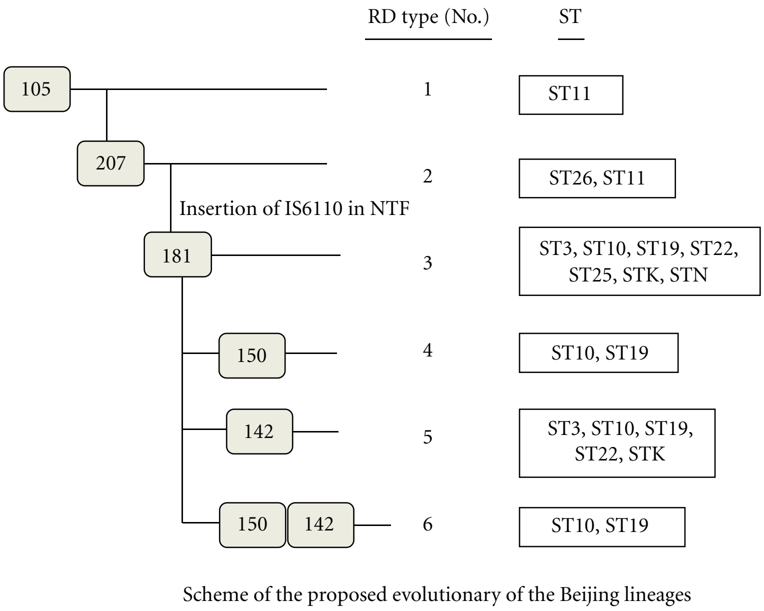

RD type (No.)
ST
105
207
Insertion of IS6110 in NTF
181
150
142
150
142
1
2
3
4
5
6
ST11
ST26, ST11
ST3, ST10, ST19, ST22, ST25, STK, STN
ST10, ST19
ST3, ST10, ST19, ST22, STK
ST10, ST19
Scheme of the proposed evolutionary of the Beijing lineages

Figure 1: Proposed origins and routes of spread of four strains of MTB to Taiwan.

mainland China due to the civil war there [4]; in total, about two million mainland Chinese have migrated to Taiwan to date. Taiwan was occupied by the Dutch beginning in 1660 for 40 years, and by the Japanese from 1895 until 1945. There are 12 tribes of aboriginals on this island, which are presumed to represent the ethnics who have inhabited the island for at least four thousand years (Figure 1).

Although both the incidence and mortality rate of TB have shown steadily declined since 1950, TB is still a leading notifiable infectious disease on Taiwan. The populations of Taiwan that have tuberculosis among them include aborigines, veterans, and Taiwanese (Hoklo). Therefore, the heterogeneous components of ethnic populations constitute a good model with which to study MTB transmission and host-pathogen relationships. An important question to be answered here is whether distinct genotypes or lineages of MTB are distributed differently according to their hosts' ethnic origins and birthplaces. In the past years, we applied MIRU-VNTR sequences as genetic markers and discovered interesting findings on the origin and evolution of MTB in Taiwan, as described below.

2. Associations of *Mycobacterium tuberculosis* Genotypes with Different Ethnic and Migratory Populations in Taiwan

Some epidemiologic studies have revealed that MTB genotype distribution is closely associated with geography, ethnicity, and population migrations [4, 5, 7]. Similar phylogeographical population structures have been reported for other human pathogens [10–13], some of which have been linked to ancient human migrations [11, 12, 14].

In Taiwan, TB is a major disease with an annual incidence of about 16,000 confirmed cases. The proportion of ethnic populations on the island is about 2% native aborigines and 98% Han Chinese (Council of Indigenous Peoples, Executive Yuan Taiwan, 2007). Previous studies in Taiwan have demonstrated a fivefold higher incidence of TB among aborigines compared to Han Chinese [15]. In addition, polymorphism at the *NRAMP1* gene appears to be associated with susceptibility to TB among aborigines but not among the Han Chinese population [15]. Preliminary studies on Beijing family MTB strains reveal differential distributions by geographic region in Taiwan [16]. These multifactor influences, including waves of immigration, allow us to trace the evolutionary history of pulmonary TB in Taiwan. Accordingly, we investigated TB evolution or transmission in (1) the aborigines of Austronesian ethnicity, whose ancestors came to Taiwan more than 500 years ago; (2) the veterans of Han Chinese origin, first-generation immigrants who moved to Taiwan 55–60 years ago; (3) the general Taiwanese population of Han Chinese, most of whose ancestors migrated to Taiwan around 200–400 years ago [4].

Based on spoligotyping classification, six distinct clades of MTB isolates among three Taiwanese subpopulations were identified: Beijing, Haarlem, East-African Indian (EAI), Latin American and Mediterranean (LAM), U, and the ill-defined T clade. Of the six known clades, the Beijing genotype overall was the most prevalent, being found in 40% of TB-positive aborigines, 72% of TB-positive veterans, and 56% of the TB-positive general population [4]. This result coincides with the global situation, with the most prevalent MTB strain worldwide being the Beijing genotype. Because Beijing strains are rapidly spreading worldwide, major TB outbreaks are most often associated with this strain [6, 17–19]. The second most frequent clade was that of the Haarlem family, which was present in 27% of aborigines and 13% of the general population, but in only 7% of veterans [4].

The third most frequent type was the T family, which was present in 5% of aborigines, 10% of veterans, and 6% of the general population. The remaining types were, in descending order of frequency, LAM, EAI, and U [4].

The Beijing family, which has the highest prevalence in the three Taiwanese subpopulations, can be further grouped into ancestral, modern, and recent strains by NTF locus analysis and RD deletion analysis. The NTF region and RD deletion are associated with the length of time since an MTB strain emerged in the human population; thus, they can be used to estimate the relative age of Beijing family clusters. Results of NTF and RD analyses revealed that ancient Beijing strains are prevalent among the aborigines, and modern Beijing strains predominate among veterans and the general population [4]. The retention of ancient characteristics of MTB among aborigines may be due to the historical tendency of Taiwan aborigines to live separately from the general population and thus have relatively little intermingling with Han Chinese.

The Haarlem genotype is the second prevalent type of TB in Taiwan. The Haarlem strain was first isolated from a patient living in Holland [20, 21] and is found mainly in Central America, the Caribbean, Europe, and West Africa, suggesting a link between Haarlem and post-Columbus Europeans [19]. *ogt* and *mgtC* gene analyses for the Haarlem lineage demonstrated that Haarlem strains circulating among aborigines in Taiwan are wild-type strains, whereas most Haarlem strains currently isolated in Europe contain single nucleotide polymorphisms (SNPs) and are comparatively modern. These results are similar to those of the Beijing strains. Given Taiwan aborigines' geographic isolation, the first transmission or exchange of Haarlem strains between the Dutch and the aborigines in Taiwan may have occurred in the 16th century during the Dutch colonization period. The late 16th century of Ming Dynasty was also the period in which Han Chinese began to migrate from mainland China to Taiwan. Thus, the Han Chinese may have introduced Beijing ancient strains into the MTB gene pool in Taiwan at that time.

3. Molecular Epidemiology and Evolutionary Genetics of *Mycobacterium tuberculosis* in Taipei

We then turn to study the strain distribution of MTB in Taipei, which is located in northern Taiwan and is the island's capital city. The strain distribution of MTB in Taipei provides us with the transmission pattern in this metropolitan city against the background described above. The city proper occupies 272 square km and has a population of 2.6 million, with an additional 4.3 million inhabitants in the surrounding metropolitan area. The population of Taipei includes the same ethnicities as described above for the entire island: Han Chinese, veterans, and Taiwanese aborigines [22]. The prevalence of TB in large urban areas such as Taipei is complicated by the close human-to-human contacts and potential multiple sources of MTB strains from different ethnic and migratory populations.

In a molecular epidemiologic analysis undertaken to investigate the prevalence of genotypes, cluster pattern, and drug resistance of MTB isolates in metropolitan Taipei, 356 MTB isolates from patients presenting with pulmonary TB were studied; the major spoligotypes found were Beijing lineages (52.5%), followed by Haarlem lineages (13.5%) and EAI plus EAI-like lineages (11%) [1]. Based on NTF and RD analyses, as well as on drug-resistance testing, strains of the Beijing family were more likely to be modern strains and have a higher percentage of multiple drug resistance than all of the other families combined. Because Han Chinese make up almost all of the general population of Taipei City, Beijing isolates found there were overwhelmingly modern strains (96%). The predominance of the Beijing strain in Taipei city constitutes a big challenge for TB control. Another important observation was that patients infected with the Beijing family were statistically younger than those infected with other genotypes (Table 1). These results suggest a possible recent spread of the Beijing genotype among younger individuals in this area. Thus, even though Taiwan has had comprehensive BCG vaccinations for more than 40 years, the predominance of the Beijing family strain in the younger cohort in our study suggests that BCG may not adequately protect young people from the Beijing strain of MTB.

This situation warrants closer attention to control policy and suggests that a better BCG vaccine is needed.

Of the 356 strains in this study, 281 isolates (79%) were sensitive to all four of the first-line agents tested and 75 (21%) were resistant to at least one drug; 2.8% were multidrug resistant (MDR) (Table 2). Analysis of the association between MDR and genotypes (as determined by spoligotyping) showed that the Beijing genotype is more likely to be MDR than all other genotypes (Haarlem, T, EAI, others, and orphan combined) ($P = .08$, OR = 3.73, and 95% C.I. = (0.78–17.83)). The EAI family is significantly more likely to be sensitive to all drugs compared to other genotypes ($P = .02$, OR = 3.64, and 95% C.I. = (1.09–12.15)). EAI belongs to a branch in the early evolution of MTB and shows more antibiotic-sensitive properties, perhaps due to a lack of drug selection pressure. Interestingly among the orphan strains, 5% were MDR and 20% were resistant to one drug, showing a distribution similar to that of the Beijing family.

Taken together, our data summarized in Figure 2 show the evolutionary relationships within the Beijing family of strains in Taipei city. RD group 1 sublineage: 1 isolate of ST11; this isolate shows a deletion of the RD105 region. RD group 2 sublineages include ST11 and ST26; these isolates show deletion of the RD105 and RD207 regions. RD group 3 sublineages include ST3, ST10, ST19, ST22, STK, and STN; these isolates show deletion of the RD105, RD207, and RD181 regions. RD group 4 sublineages include ST10 and ST19; these isolates show deletion of the RD105, RD207, RD181, and RD150 regions. RD group 5 sublineages include ST3, ST10 ST19, and ST22; these isolates show deletion of the RD105, RD207, RD181, and RD142 regions. RD group 6 sublineages include ST10 and ST19; these isolates show deletion of the RD105, RD207, RD181, RD142, and RD150 regions. It has been suggested that insertion sequence- (IS-) mediated deletion events are an important

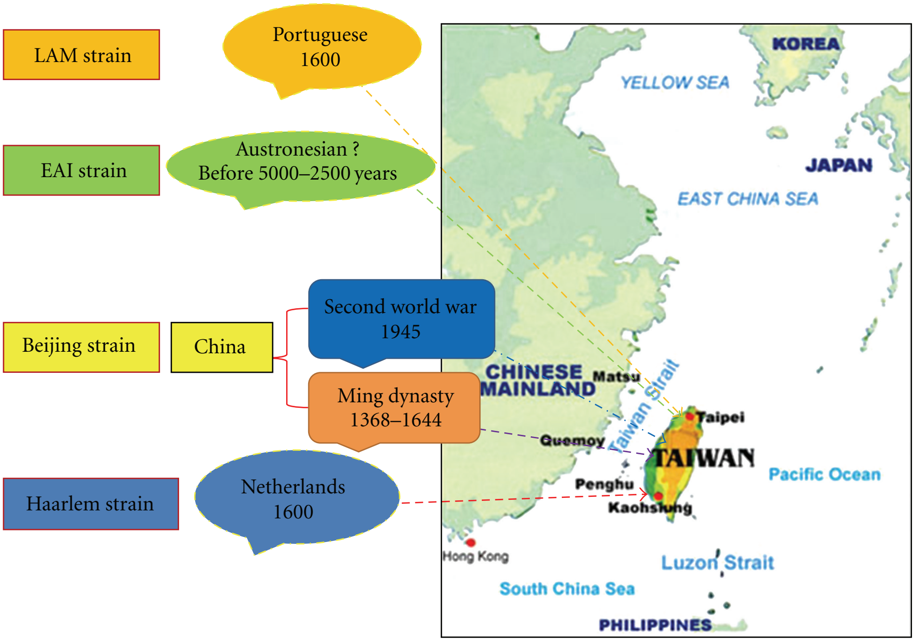


Figure 2: Scheme of the proposed evolution of Beijing lineages. The scheme is based on the deletion of genomic regions (RD: region of difference, shown in gray rectangles), and types of sequence (ST) designations from the studies of Filliol et al. [23] and Iwamoto et al. [24].

Table 1: Association of Beijing MTB genotype and different age groups of patients[a].

Age group (yr)	No. (%) isolates	No. (%) of Beijing isolates	Odds ratio	95% C.I.	*P* value
	356	187 (52.53)			
≤25	34 (9.55)	29 (85.29)	5.80	2.11–15.98	.0002
≤30	54 (15.17)	37 (68.52)	2.18	1.11–4.28	.02
31–60	95 (26.69)	50 (52.63)	1.11	0.65–1.90	.7
61–75	85 (23.88)	39 (45.88)	0.85	0.49–1.48	.56
≥76	122 (34.27)	61 (50.00)	1	reference group	

[a] Adapted from [1].

Table 2: Association between MTB genotype and drug resistance in patients[e].

Genotype family	No. of isolates (%)	MDR (%)	Any one drug (%)	All sensitivity (%)
Beijing[a]	187 (52.5)	8(4.2)	36 (19.4)	143 (76.4)
Haarlem	48 (13.5)	0	9 (18.8)	39 (81.2)
EAI[b]	40 (11.2)	0	3 (7.5)	37 (92.5)
T	25 (7.1)	0	8 (32.0)	17 (68.0)
"Others"[c] (LAM, U, MANU, Bovis1)	16 (4.5)	0	1 (6.3)	15 (93.7)
Unclassified[d]	40 (11.2)	2 (5)	8 (20)	30 (75)
Total	356	10 (2.8)	65 (18.2)	281 (79)

[a] Including Beijing-like strains.
[b] Including EAI-like strains.
[c] "Others", all genotype families with a frequency of less than 10 cases.
[d] Unclassified, no internationally recognized genotype family assigned, based on the SpolDB4 spoligotype database.
[e] Adapted from [1].

mechanism driving mycobacterial genome variation. Based on our results (Figure 2), the RD105 and RD207 deletions appear to have been early events in the evolutionary history of Beijing strains; however, the IS6110 insertion occurred after the RD181 deletion but has not always persisted in later sublineage evolution. Thus, neither of the RD type 1 and type 2 groups (which include ST26 and ST11) have an IS6110 insert in the NTF region (N family). We still found some characteristics of ancient Beijing strains (N family) in ST19, ST10, and ST22.

Figure 1 illustrates the proposed origins and routes of spread of four strains of MTB in Taiwan.

Route 1. The Beijing strain may have migrated to Taiwan through two separate historic events: the first during the Ming dynasty and the second wave shortly after World War II. Through these two migrations, the ancient Beijing strain has evolved into the modern Beijing strain.

Route 2. Haarlem originated in the Netherlands. It migrated to Taiwan during the Dutch reign over the island in the 16th century and continues to be a major strain here. It is also important to note that there has been no observed genetic mutation in the strain that was passed onto the natives of Taiwan. The Haarlem strain that remained in the Netherlands, however, has mutations in the *ogt* and *mgtC* genes, thus, resulting in SNP variants.

Route 3. LAM originated in both Europe and the Americas. It may have migrated to Taiwan during the Portuguese reign in the 16th century and been passed on to the natives of Taiwan.

Route 4. EAI originated in Taiwanese aborigines, entering Taiwan four thousand years ago. It may be closely associated with the Austronesian culture. The Austronesian peoples are a population in Oceania and Southeast Asia who speak languages of the Austronesian family. They include Taiwanese aborigines; the majority ethnic groups of East Timor, Indonesia, Malaysia, the Philippines, Brunei, Madagascar, Micronesia, and Polynesia; the Polynesian peoples of New Zealand and Hawaii and the Austronesian peoples of Melanesia.

Problems are remaining to be solved. Molecular genetic analysis of clinical MTB strains delineates relationships among closely related strains of pathogenic microbes and allows construction of genetic frameworks for examining the distribution of biomedically relevant traits such as virulence, transmissibility, and host range. Based on the strain distribution in different ethnic populations, we will attempt to identify factors that determine the disease transmission. Comparative genomic hybridization (CGH) microarray chips will be designed based on the genomic sequence to conduct the population genetic study efficiently. The information we provided in this paper will help us to better understand the dynamics of TB transmission in Taiwan and hence is a good model to understand the global distribution of MTB strains among different geographic regions and ethnic populations.

Conflict of Interests

The authors declare no conflict of interests.

Acknowledgments

This project was supported by grants from the National Health Research Institutes. All participants of this consortium are acknowledged for valuable discussions.

References

[1] H. Y. Dou, F. C. Tseng, C. W. Lin et al., "Molecular epidemiology and evolutionary genetics of Mycobacterium tuberculosis in Taipei," *BMC Infectious Diseases*, vol. 8, article 170, 2008.

[2] World Health Oraganization, "Global Tuberculosis Control. Surveillance, Planning, Financing," 2007.

[3] B. Mathema, N. E. Kurepina, P. J. Bifani, and B. N. Kreiswirth, "Molecular epidemiology of tuberculosis: current insights," *Clinical Microbiology Reviews*, vol. 19, no. 4, pp. 658–685, 2006.

[4] H. Y. Dou, F. C. Tseng, J. J. Lu et al., "Associations of Mycobacterium tuberculosis genotypes with different ethnic and migratory populations in Taiwan," *Infection, Genetics and Evolution*, vol. 8, no. 3, pp. 323–330, 2008.

[5] S. Gagneux, K. DeRiemer, T. Van et al., "Variable host-pathogen compatibility in Mycobacterium tuberculosis," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, no. 8, pp. 2869–2873, 2006.

[6] J. R. Glynn, J. Whiteley, P. J. Bifani, K. Kremer, and D. van Soolingen, "Worldwide occurrence of Beijing/W strains of Mycobacterium tuberculosis: a systematic review," *Emerging Infectious Diseases*, vol. 8, no. 8, pp. 843–849, 2002.

[7] A. E. Hirsh, A. G. Tsolaki, K. DeRiemer, M. W. Feldman, and P. M. Small, "Stable association between strains of Mycobacterium tuberculosis and their human host populations," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 14, pp. 4871–4876, 2004.

[8] P. Supply, R. M. Warren, A. L. Bañuls et al., "Linkage disequilibrium between minisatellite loci supports clonal evolution of Mycobacterium tuberculosis in a high tuberculosis incidence area," *Molecular Microbiology*, vol. 47, no. 2, pp. 529–538, 2003.

[9] P. Supply, E. Mazars, S. Lesjean, V. Vincent, B. Gicquel, and C. Locht, "Variable human minisatellite-like regions in the Mycobacterium tuberculosis genome," *Molecular Microbiology*, vol. 36, no. 3, pp. 762–771, 2000.

[10] H. T. Agostini, R. Yanagihara, V. Davis, C. F. Ryschkewitsch, and G. L. Stoner, "Asian genotypes of JC virus in Native Americans and in a Pacific Island population: markers of viral evolution and human migration," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 94, no. 26, pp. 14542–14546, 1997.

[11] D. Falush, T. Wirth, B. Linz et al., "Traces of human migrations in Helicobacter pylori populations," *Science*, vol. 299, no. 5612, pp. 1582–1585, 2003.

[12] M. Monot, N. Honoré, T. Garnier et al., "On the origin of leprosy," *Science*, vol. 308, no. 5724, pp. 1040–1042, 2005.

[13] J. M. Musser, J. S. Kroll, D. M. Granoff et al., "Global genetic structure and molecular epidemiology of encapsulated Haemophilus influenzae," *Reviews of Infectious Diseases*, vol. 12, no. 1, pp. 75–111, 1990.

[14] T. Wirth, X. Wang, B. Linz et al., "Distinguishing human ethnic groups by means of sequences from Helicobacter pylori: lessons from Ladakh," *Proceedings of the National Academy of*

Sciences of the United States of America, vol. 101, no. 14, pp. 4746–4751, 2004.

[15] Y. H. Hsu, C. W. Chen, H. S. Sun, R. Jou, J. J. Lee, and I. J. Su, "Association of NRAMP 1 gene polymorphism with susceptibility to tuberculosis in Taiwanese aboriginals," *Journal of the Formosan Medical Association*, vol. 105, no. 5, pp. 363–369, 2006.

[16] R. Jou, C. Y. Chiang, and W. L. Huang, "Distribution of the Beijing family genotypes of Mycobacterium tuberculosis in Taiwan," *Journal of Clinical Microbiology*, vol. 43, no. 1, pp. 95–100, 2005.

[17] J. R. Glynn, K. Kremer, M. W. Borgdorff, M. P. Rodriguez, and D. van Soolingen, "Beijing/W genotype Mycobacterium tuberculosis and drug resistance: European concerted action on new generation genetic markers and techniques for the epidemiology and control of tuberculosis," *Emerging Infectious Diseases*, vol. 12, no. 5, pp. 736–743, 2006.

[18] P. J. Bifani, B. Mathema, N. E. Kurepina, and B. N. Kreiswirth, "Global dissemination of the Mycobacterium tuberculosis W-Beijing family strains," *Trends in Microbiology*, vol. 10, no. 1, pp. 45–52, 2002.

[19] K. Brudey, J. R. Driscoll, L. Rigouts et al., "Mycobacterium tuberculosis complex genetic diversity: mining the fourth international spoligotyping database (SpolDB4) for classification, population genetics and epidemiology," *BMC Microbiology*, vol. 6, article 23, 2006.

[20] P. Farnia, M. R. Masjedi, M. Mirsaeidi et al., "Prevalence of Haarlem I and Beijing types of Mycobacterium tuberculosis strains in Iranian and Afghan MDR-TB patients," *Journal of Infection*, vol. 53, no. 5, pp. 331–336, 2006.

[21] K. Kremer, D. van Soolingen, R. Frothingham et al., "Comparison of methods based on different molecular epidemiological markers for typing of Mycobacterium tuberculosis complex strains: Interlaboratory study of discriminatory power and reproducibility," *Journal of Clinical Microbiology*, vol. 37, no. 8, pp. 2607–2618, 1999.

[22] "A Brief History of Taiwan—A Sparrow Transformed into a Phoenix," http://www.gio.gov.tw/Taiwan-Website/5-gp/history/.

[23] I. Filliol, J. R. Driscoll, D. van Soolingen et al., "Snapshot of moving and expanding clones of Mycobacterium tuberculosis and their global distribution assessed by spoligotyping in an international study," *Journal of Clinical Microbiology*, vol. 41, no. 5, pp. 1963–1970, 2003.

[24] T. Iwamoto, S. Yoshida, K. Suzuki, and T. Wada, "Population structure analysis of the Mycobacterium tuberculosis Beijing family indicates an association between certain sublineages and multidrug resistance," *Antimicrobial Agents and Chemotherapy*, vol. 52, no. 10, pp. 3805–3809, 2008.

SAGE-Hindawi Access to Research
International Journal of Evolutionary Biology
Volume 2011, Article ID 685015, 30 pages
doi:10.4061/2011/685015

Research Article

Distribution of Genes Encoding Nucleoid-Associated Protein Homologs in Plasmids

Toshiharu Takeda,[1] Choong-Soo Yun,[1,2] Masaki Shintani,[3] Hisakazu Yamane,[1] and Hideaki Nojiri[1,2]

[1] *Biotechnology Research Center, The University of Tokyo, 1-1-1 Yayoi, Bunkyo-ku, Tokyo 113-8657, Japan*
[2] *Agricultural Bioinformatics Research Unit, The University of Tokyo, 1-1-1 Yayoi, Bunkyo-ku, Tokyo 113-8657, Japan*
[3] *Japan Collection of Microorganisms, RIKEN Bioresource Center, 2-1 Hirosawa, Wako, Saitama 351-0198, Japan*

Correspondence should be addressed to Hideaki Nojiri, anojiri@mail.ecc.u-tokyo.ac.jp

Received 14 October 2010; Accepted 27 November 2010

Academic Editor: Hiromi Nishida

Bacterial nucleoid-associated proteins (NAPs) form nucleoprotein complexes and influence the expression of genes. Recent studies have shown that some plasmids carry genes encoding NAP homologs, which play important roles in transcriptional regulation networks between plasmids and host chromosomes. In this study, we determined the distributions of the well-known NAPs Fis, H-NS, HU, IHF, and Lrp and the newly found NAPs MvaT and NdpA among the whole-sequenced 1382 plasmids found in Gram-negative bacteria. Comparisons between NAP distributions and plasmid features (size, G+C content, and putative transferability) were also performed. We found that larger plasmids frequently have NAP gene homologs. Plasmids with H-NS gene homologs had less G+C content. It should be noted that plasmids with the NAP gene homolog also carried the relaxase gene involved in the conjugative transfer of plasmids more frequently than did those without the NAP gene homolog, implying that plasmid-encoded NAP homologs positively contribute to transmissible plasmids.

1. Introduction

Bacterial chromosomal DNA is folded to form a compacted structure, the nucleoid. The proteins involved in folding the chromosome are known as nucleoid-associated proteins (NAPs) [1, 2]. Because of their DNA-binding ability, NAPs can also play an important role in global gene regulation [1, 2]. Each well-known NAP in *Enterobacteriaceae* may be categorized as a "factor for inversion stimulation" (Fis), "histone-like nucleoid structuring protein" (H-NS), "histone-like protein from *Escherichia coli* strain U93" (HU), "integration host factor" (IHF), or "leucine-responsive regulatory protein" (Lrp) [1]. Fis is one of the most abundant NAPs in exponentially growing *E. coli* cells, and its role as a transcriptional regulator has been investigated [3]. H-NS binds DNA, especially A+T-rich regions including promoter regions or horizontally acquired DNA and acts as a global transcriptional repressor [4]. HU and IHF are similar in amino acid sequence level, and both are global regulators [5, 6], although they have distinct DNA-binding activities: HU binds to DNA nonspecifically whereas IHF binds to a consensus sequence [7]. Lrp has a global influence on transcription regulation and is also involved in microbial virulence [8]. In addition to these well-known NAPs, many other NAPs are found not only in *Enterobacteriaceae* but also in other organisms. For instance, NdpA, a functionally unknown NAP, has been found in Gram-negative bacteria [9]. The MvaT family protein is the functional homolog of H-NS in *Pseudomonas* bacteria [10].

Horizontal gene transfer (HGT), which is mediated by transduction, transformation, and conjugation, plays an important role in the evolution of prokaryotic genomes [11, 12]. Genes acquired by HGT can provide beneficial functions such as resistance to antibiotics and advantages to their host under selective pressures [13]. However, the mechanisms underlying the integration of newly acquired genes into host regulatory networks are still unclear. Recent investigations have shown that some plasmids carry the genes encoding NAP homologs, which play important roles in transcriptional regulation networks between plasmids and host chromosomes and in maintaining host cell fitness. For example, Doyle et al. [14] reported that plasmid-encoded H-NS-like protein has a "stealth" function that allows for plasmid transfer into host cells without disrupting host regulatory networks, maintaining host cell fitness. Yun and Suzuki et al. [15] reported that plasmid-encoded H-NS-like protein can also play a key role in optimizing gene transcription both on the plasmid and in the host chromosome.

In this study, we determined the distributions of NAP homologs among plasmids and discussed their roles in the maintenance of plasmid and host cell fitness.

2. Materials and Methods

2.1. Plasmid Database Collection and Local BLAST Analyses. The completely sequenced plasmid database was downloaded from the NCBI ftp site (ftp://ftp.ncbi.nih.gov/genomes/Plasmids/). Some duplicated sequence data of the same plasmids were removed from the database. Identification of plasmids that contain the genes encoding NAP homologs was performed using the local TBLASTN program (ver. 2.2.24, ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/) under strict conditions (i.e., thresholds of 30% identity and 70% query coverage). The complete amino acid sequences of Fis (DDBJ/EMBL/GenBank accession no. AP_003801), H-NS (AP_001863), Hha (AP_001109), HUα (AP_003818), HUβ (AP_001090), IHFα (AP_002332), IHFβ (AP_001542), Lrp (AP_001519), and NdpA (P33920) from *E. coli* K-12 W3110 and MvaT (AAP33788) from *Pseudomonas aeruginosa* PAO1 were used as query sequences.

2.2. Bacterial Genome Analyses. The complete genome sequences of bacteria were downloaded from the NCBI ftp site (ftp://ftp.ncbi.nih.gov/genomes/Bacteria/). The number of NAP genes on proteobacterial genomes was investigated using the TBLASTN program (http://www.ncbi.nlm.nih.gov/sutils/genom_table.cgi) under strict conditions (i.e., thresholds of 30% identity and 70% query coverage).

2.3. Plasmid Classification. Plasmids in the database were classified into six groups according to their source organisms: Gram-negative, Gram-positive, archaeal, eukaryotic, viral, and unclassified. Putative transferability of each Gram-negative plasmid was determined by whether it carried the relaxase gene of each MOB family that Garcillán-Barcia et al. proposed [16]. Instead of using the local PSI-BLAST program (ver. 2.2.24, ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/) as described by Garcillán-Barcia et al. [16], we used the local TBLASTN program.

3. Results and Discussion

3.1. Database Collection and Plasmid Classification by Origin. We downloaded the whole sequences of 2278 plasmids from the NCBI ftp site (April 2010). Duplicated plasmids were removed manually, and the resultant 2260 plasmid sequences were used in this study. To understand what types of plasmids were included in the database, we classified them into six groups according to their source organisms. The database included 1382 Gram-negative, 725 Gram-positive, 81 archaeal, 43 eukaryotic, 1 viral, and 28 unclassified plasmids.

3.2. Identification of the Plasmids Containing NAP Gene Homologs. Using the amino acid sequences of well-known NAPs (Fis, H-NS, HU, IHF, and Lrp) and newly found NAPs (MvaT and NdpA), their distributions were surveyed for plasmids using the TBLASTN program. Some plasmids had ORFs showing sequence similarities to both HU and IHF. We adopted the one with the higher E value. Of 2260 plasmids, 155 (7%) contained the gene encoding NAP homolog. Of those, 116 (75%) contained only one NAP gene homolog and 39 (25%) contained more than one NAP gene homolog. No plasmids carried the Fis gene homolog. Twenty-two plasmids carried the H-NS gene homolog, and all of them had a Gram-negative origin (Table 1). Sixty-six plasmids had the HU gene homolog; of these, 51 had a Gram-negative origin and 15 had a Gram-positive origin (Table 2). Twenty-seven plasmids (25 with Gram-negative and 2 with Gram-positive origins) carried the IHF gene homolog (Table 3). Forty-eight plasmids (46 with Gram-negative, 1 with a Gram-positive, and 1 with an archaeal origin) carried the Lrp gene homolog (Table 4). Of these, 23 (48%) contained more than one Lrp gene homolog. On the other hand, MvaT and NdpA homologs were encoded on only 3 plasmids, and all of them were of Gram-negative origin (Table 5). Previously reported plasmids that are known to have NAP gene homologs were included in those 155 plasmids. These included R27 (NC_002305) and pHCM1 (NC_003384) [18, 19] with the H-NS gene homolog; pQBR103 (NC_009444) [20] with the HU and NdpA gene homologs; and pCAR1 (NC_004444) [21, 22] with the MvaT, HU, and NdpA gene homologs. These results indicated the adequacy of our search. Because we used NAPs from Gram-negative bacteria as query sequences, it may be reasonable that 136 (88%) of 155 plasmids with the NAP gene homolog belonged to the group isolated from Gram-negative bacteria. Therefore, in further studies we discussed the Gram-negative plasmid group.

3.3. Relationships between Plasmid Size and NAP Gene Homolog Distributions. We first compared the sizes of 136 plasmids with NAP gene homologs with those of all 1382 Gram-negative group plasmids. All 1382 plasmids could be divided into 4 groups according to size, small (<10 kb), intermediate (10 to 100 kb), large (100 kb to 1 Mb), and mega (>1 Mb) plasmids. The distribution of the 136 plasmids, each of which had one or more genes encoding NAP homologs, is shown in Figure 1(a): none of 415 small plasmids, 34 (5%) of 686 intermediate plasmids, 90 (33%) of 269 large plasmids, and 12 (100%) of 12 mega plasmids carried at least one NAP gene homolog. The average size of the 136 plasmids was larger (364 kb) than that of all 1382 plasmids

Table 1: Plasmids containing the gene encoding H-NS homolog[a].

Plasmid name	Accession no.	Source organism	Length (nt)	G+C content (%)[b]	Identity (%)[c]	Query coverage (%)	Subject start	Subject end	Classification[d]	MOB family[e]
1	NC_013972	*Erwinia amylovora* ATCC 49946	28243	50	66	99	3129	2728	–	
pAsa5	NC_009350	*Aeromonas salmonicida* subsp. *salmonicida* A449	155098	54	46	99	941	534	–	MOB_F
					47	99	16890	16483		
pAsal5	NC_009352	*Aeromonas salmonicida* subsp. *salmonicida*	18536	54	46	99	12285	12692	–	
pEA29	NC_013957	*Erwinia amylovora* CFBP1430	28259	50	66	99	3129	2728	–	
pEA29	NC_005706	*Erwinia amylovora*	28185	50	64	99	2991	2590	–	
pEC-IMP	NC_012555	*Enterobacter cloacae*	318782	48	64	99	109370	108969	–	MOB_H
pEC-IMPQ	NC_012556	*Enterobacter cloacae*	324503	48	64	99	109370	108969	–	MOB_H
pEJ30	NC_004834	*Erwinia* sp. Ejp 556	29593	50	66	99	4651	4250	–	
pEP36	NC_013263	*Erwinia pyrifoliae* Ep1/96	35909	50	66	99	25040	25441	–	
pEP36	NC_004445	*Erwinia pyrifoliae* Ep1/96	35904	50	64	98	4675	4280	–	
pET45	NC_010699	*Erwinia tasmaniensis* Et1/99	44694	51	52	93	37435	37809	–	MOB_F
pET49	NC_010697	*Erwinia tasmaniensis* Et1/99	48751	44	36	94	30821	31204	–	
pHCM1	NC_003384	*Salmonella enterica* subsp. *enterica* serovar Typhi str. CT18	218160	48	61	99	131861	131460	–	MOB_H
pK2044	NC_006625	*Klebsiella pneumoniae* NTUH-K2044	224152	50	67	99	35717	36112	–	
plasmid_153kb	NC_009705	Yersinia pseudotuberculosis IP 31758	153140	40	44	100	139846	140265	–	
pLVPK	NC_005249	*Klebsiella pneumoniae*	219385	50	67	99	114397	114792	–	
pMAK1	NC_009981	*Salmonella enterica* subsp. *enterica* serovar Choleraesuis	208409	47	61	99	60046	59645	–	MOB_H

Table 1: Continued.

Plasmid name	Accession no.	Source organism	Length (nt)	G+C content (%)[b]	Identity (%)[c]	Query coverage (%)	Subject start	Subject end	Classification[d]	MOB family[e]
pO111_1	NC_013365	*Escherichia coli* O111:H- str. 11128	204604	47	61	99	80175	79774	–	MOB_H
pSG1	NC_007713	*Sodalis glossinidius* str. "*morsitans*"	83306	49	43	97	2533	2922	–	
R27	NC_002305	*Salmonella enterica* subsp. *enterica* serovar Typhi	180461	46	61	99	148225	148626	–	MOB_H
R478	NC_005211	*Serratia marcescens*	274762	46	64	99	111747	111346	–	MOB_H
Unnamed	NC_011148	*Salmonella enterica* subsp. *enterica* serovar Agona str. SL483	37978	41	43	95	7671	7288	–	

[a] This list is the result of a TBLASTN analysis using the amino acid sequence of H-NS as a query under strict conditions (i.e., thresholds of 30% identity and 70% query coverage). Besides these plasmids, pSf-R27 from *Shigella flexneri* 2a str. 2457T was completely sequenced by Wei et al. [17] and encodes the H-NS-like protein Sfh.

[b] Average G+C content of the plasmid.

[c] Reported TBLASTN identity to H-NS.

[d] Plasmid classification according to its source organism (–, Gram-negative plasmid).

[e] Plasmid classification according to its relaxase gene sequence as described by Garcillán-Barcia et al. [16].

Table 2: Plasmids containing the gene encoding HU homolog[a].

Plasmid name	Accession no.	Source organism	Length (nt)	G+C content (%)[b]	Identity (%)[c]	Query coverage (%)	Subject start	Subject end	Classification[d]	MOB family[e]
1	NC_006823	*Aromatoleum aromaticum* EbN1	207355	58	55	99	186175	185909	–	
1	NC_007949	*Polaromonas* sp. JS666	360405	57	52	99	61052	60786	–	MOB$_{\text{H}}$
1	NC_008010	*Deinococcus geothermalis* DSM 11300	574127	66	38	97	550805	550545	+	
1	NC_008503	*Lactococcus lactis* subsp. *cremoris* SK11	14041	34	37	94	9732	10007	+	MOB$_{\text{P}}$
1	NC_008242	*Chelativorans* sp. BNC1	343931	62	41	94	133932	133678	–	MOB$_{\text{Q}}$
2	NC_012529	*Deinococcus deserti* VCD115	314317	64	38	93	269648	269899	+	
3	NC_012528	*Deinococcus deserti* VCD115	396459	61	40	96	8700	8957	+	
Megaplasmid	NC_007974	*Cupriavidus metallidurans* CH34	2580084	64	51	99	1393415	1393149	–	MOB$_{\text{V}}$
Megaplasmid	NC_005863	*Desulfovibrio vulgaris* str. *Hildenborough*	202301	66	31	98	5502	5765	–	
Megaplasmid pDF308	NC_013940	*Deferribacter desulfuricans* SSM1	308544	24	41	100	253817	253548	–	
Megaplasmid pHG1	NC_005241	*Ralstonia eutropha* H16	452156	62	48	99	343060	342791	–	
p49879.1	NC_006907	*Leptospirillum ferrooxidans*	28878	58	47	99	3281	3015	–	MOB$_{\text{Q}}$
p49879.2	NC_006909	*Leptospirillum ferrooxidans*	28012	55	48	99	15858	15592	–	MOB$_{\text{Q}}$
pAH187_270	NC_011655	*Bacillus cereus* AH187	270082	34	59	100	113139	112870	+	
pAH820_272	NC_011777	*Bacillus cereus* AH820	272145	34	58	100	153060	152791	+	
pAM04528	NC_012693	*Salmonella enterica*	158213	52	57	99	14067	14333	–	MOB$_{\text{H}}$
pAOVO01	NC_008765	*Acidovorax* sp. JS42	72689	62	46	100	65140	64871	–	MOB$_{\text{F}}$
pAPA01-011	NC_013210	*Acetobacter pasteurianus* IFO 3283-01	191799	53	47	100	154736	154467	–	
					46	99	38442	38708		
pAR060302	NC_012692	*Escherichia coli*	166530	53	57	99	15755	16021	–	MOB$_{\text{H}}$
pAsa4	NC_009349	*Aeromonas salmonicida* subsp. *salmonicida* A449	166749	53	60	99	26844	26578	–	MOB$_{\text{H}}$
pAtS4c	NC_011984	*Agrobacterium vitis* S4	211620	59	45	94	141245	140991	–	MOB$_{\text{Q}}$
pAtS4e	NC_011981	*Agrobacterium vitis* S4	631775	57	41	94	40476	40222	–	MOB$_{\text{Q}}$
pBc239	NC_011973	*Bacillus cereus* Q1	239246	33	52	100	191895	192164	+	
pBF9343	NC_006873	*Bacteroides fragilis* NCTC 9343	36560	32	35	92	15803	15558	–	MOB$_{\text{P}}$
pBPHY01	NC_010625	*Burkholderia phymatum* STM815	1904893	62	43	99	826527	826252	–	

Table 2: Continued.

Plasmid name	Accession no.	Source organism	Length (nt)	G+C content (%)[b]	Identity (%)[c]	Query coverage (%)	Subject start	Subject end	Classification[d]	MOB family[e]
pBPHY02	NC_010627	*Burkholderia phymatum* STM815	595108	59	45	99	98625	98359	–	
pBtoxis	NC_010076	*Bacillus thuringiensis* serovar israelensis	127923	32	52	99	77382	77648	+	
pBWB401	NC_010180	*Bacillus weihenstephanensis* KBAB4	417054	34	59	100	338347	338078	+	
pCAR1	NC_004444	*Pseudomonas resinovorans*	199035	56	42	99	97809	98075	–	MOB_H
pCAUL01	NC_010335	*Caulobacter* sp. K31	233649	67	44	99	97598	97329	–	MOB_Q
pCER270	NC_010924	*Bacillus cereus*	270082	34	59	100	169548	169279	+	
pDBORO	NC_009137	*Lactococcus lactis* subsp. *lactis* bv. diacetylactis	16404	35	37	94	16387	16112	+	
pDVUL01	NC_008741	*Desulfovibrio vulgaris* DP4	198504	66	31	98	198317	198054	–	
peH4H	NC_012690	*Escherichia coli*	148105	53	57	99	14067	14333	–	MOB_H
pG9842_209	NC_011775	*Bacillus cereus* G9842	209488	30	60	100	88828	88559	+	
pH308197_258	NC_011339	*Bacillus cereus* H3081.97	258484	34	59	100	83033	83302	+	
pHD5AT	NC_012752	*Candidatus Hamiltonella defensa* 5AT (*Acyrthosiphon pisum*)	59032	45	45	99	14981	15247	–	MOB_P
pIP1202	NC_009141	*Yersinia pestis* bv. Orientalis str. IP275	182913	53	57	99	14067	14333	–	MOB_H
plasmid 2	NC_007972	*Cupriavidus metallidurans* CH34	171459	61	46	99	125530	125261	–	
pMOL28	NC_006525	*Cupriavidus metallidurans* CH34	171461	61	46	99	51529	51798	–	
pMP118	NC_007930	*Lactobacillus salivarius* UCC118	242436	32	54	99	56763	56497	+	MOB_V
pNPUN02	NC_010632	*Nostoc punctiforme* PCC 73102	254918	41	44	99	74804	74538	–	MOB_V
pOANT02	NC_009670	*Ochrobactrum anthropi* ATCC 49188	101491	59	49	94	32700	32446	–	
pP91278	NC_008613	*Photobacterium damselae* subsp. *piscicida*	131520	52	57	99	125918	126184	–	MOB_H
pP99-018	NC_008612	*Photobacterium damselae* subsp. *piscicida*	150157	51	57	99	133314	133580	–	MOB_H
pPER272	NC_010921	*Bacillus cereus*	272145	34	58	100	153060	152791	+	
pPMA4326A	NC_005918	*Pseudomonas syringae* pv. *maculicola*	46697	55	42	99	1520	1786	–	
pPMA4326B	NC_005919	*Pseudomonas syringae* pv. *maculicola*	40110	55	45	99	1457	1723	–	

Table 2: Continued.

Plasmid name	Accession no.	Source organism	Length (nt)	G+C content (%)[b]	Identity (%)[c]	Query coverage (%)	Subject start	Subject end	Classification[d]	MOB family[e]
pQBR103	NC_009444	*Pseudomonas fluorescens* SBW25	425094	53	51	99	182862	183128	–	
pR132503	NC_012853	*Rhizobium leguminosarum* bv. trifolii WSM1325	516088	59	47	94	300662	300916	–	MOB_Q
pRA1	NC_012885	*Aeromonas hydrophila*	143963	51	58	99	15573	15839	–	MOB_H
pRALTA	NC_010529	*Cupriavidus taiwanensis*	557200	60	46	98	153542	153276	–	
pREB1	NC_009926	*Acaryochloris marina* MBIC11017	374161	47	46	100	339743	340012	–	MOB_F
pREB2	NC_009927	*Acaryochloris marina* MBIC11017	356087	45	48	100	57583	57852	–	MOB_F
pREB3	NC_009928	*Acaryochloris marina* MBIC11017	273121	45	46	100	234682	234951	–	MOB_F
					42	100	243339	243608		
pRL7	NC_008382	*Rhizobium leguminosarum* bv. viciae 3841	151564	58	48	94	20484	20230	–	MOB_Q
pRLG203	NC_011370	*Rhizobium leguminosarum* bv. trifolii WSM2304	308747	58	49	94	141121	140867	–	
pRp12D01	NC_012855	*Ralstonia pickettii* 12D	389779	58	37	99	321346	321080	–	MOB_H
pSG2	NC_007184	*Sodalis glossinidius*	27240	45	45	86	10072	9845	–	
pSG3	NC_007186	*Sodalis glossinidius*	19201	51	51	100	13812	13543	–	
pSN254	NC_009140	*Salmonella enterica* subsp. *enterica* serovar Newport str. SL254	176473	53	57	99	14067	14333	–	MOB_H
pTiS4	NC_011982	*Agrobacterium vitis* S4	258824	57	41	94	27356	27102	–	MOB_Q
					40	94	83408	83154		
pTi-SAKURA	NC_002147	*Agrobacterium tumefaciens*	206479	56	44	94	95763	95509	–	MOB_Q
pVSAL840	NC_011311	*Aliivibrio salmonicida* LFI1238	83540	40	60	99	31361	31627	–	MOB_F
					58	99	77350	77084		
pYR1	NC_009139	*Yersinia ruckeri*	158038	51	57	99	15070	15336	–	MOB_H
Ti	NC_003065	*Agrobacterium tumefaciens* str. C58	214233	57	44	94	139735	139481	–	MOB_Q

[a] This list is the result of a TBLASTN analysis using the amino acid sequence of HUα or HUβ as a query under strict conditions (i.e., thresholds of 30% identity and 70% query coverage).

[b] Average G+C content of the plasmid.

[c] Reported TBLASTN identity to HU.

[d] Plasmid classification according to its source organism (−, Gram-negative plasmid; +, Gram-positive plasmid).

[e] Plasmid classification according to its relaxase gene sequence as described by Garcillán-Barcia et al. [16].

Table 3: Plasmids containing the gene encoding IHF homolog[a].

Plasmid name	Accession no.	Source organism	Length (nt)	G+C content (%)[b]	Identity (%)[c]	Query coverage (%)	Subject start	Subject end	Classification[d]	MOB family[e]
At	NC_003064	*Agrobacterium tumefaciens* str. C58	542868	57	36	82	112654	112412	–	MOB_Q
Megaplasmid	NC_012811	*Methylobacterium extorquens* AM1	1261460	68	33	94	720582	720860	–	
p2META1	NC_012809	*Methylobacterium extorquens* AM1	37858	65	44	95	28369	28635	–	MOB_Q
pAACI01	NC_013206	*Alicyclobacillus acidocaldarius* subsp. *acidocaldarius* DSM 446	91726	54	43	80	62668	62432	+	
pACHL01	NC_011879	*Arthrobacter chlorophenolicus* A6	426858	64	32	92	408818	408546	+	
pALVIN02	NC_013862	*Allochromatium vinosum* DSM 180	39929	53	60	98	10902	10627	–	
pAph01	NC_013193	*Candidatus Accumulibacter phosphatis clade IIA* str. UW-1	167595	62	56	95	144197	144463	–	MOB_P
pAph03	NC_013191	*Candidatus Accumulibacter phosphatis clade IIA* str. UW-1	37695	59	58	97	5412	5140	–	
pAtK84b	NC_011990	*Agrobacterium radiobacter* K84	184668	59	38	86	54109	53855	–	MOB_Q
pAtK84c	NC_011987	*Agrobacterium radiobacter* K84	388169	57	43	93	340807	340532	–	
					46	93	10327	10052		
pAtS4b	NC_011991	*Agrobacterium vitis* S4	130435	56	47	97	44880	45152	–	MOB_Q
pBBta01	NC_009475	*Bradyrhizobium* sp. BTAi1	228826	61	39	86	6642	6388	–	
pBFY46	NC_006297	*Bacteroides fragilis* YCH46	33716	34	35	89	25098	25343	–	MOB_P
pBIND01	NC_010580	*Beijerinckia indica* subsp. *indica* ATCC 9039	181736	56	36	77	179816	179601	–	MOB_F
pCHQ1	NC_014007	*Sphingobium japonicum* UT26S	190974	63	36	90	63111	63377	–	
pGLOV01	NC_010815	*Geobacter lovleyi* SZ	77113	53	38	92	41196	41468	–	
pM44601	NC_010373	*Methylobacterium* sp. 4-46	57951	65	35	97	7806	7534	–	
pMPOP01	NC_010727	*Methylobacterium populi* BJ001	25164	65	49	93	10635	10375	–	
pMRAD03	NC_010514	*Methylobacterium radiotolerans* JCM 2831	42985	63	38	94	26778	26515	–	MOB_F

Table 3: Continued.

Plasmid name	Accession no.	Source organism	Length (nt)	G+C content (%)[b]	Identity (%)[c]	Query coverage (%)	Subject start	Subject end	Classification[d]	MOB family[e]
pMRAD04	NC_010517	*Methylobacterium radiotolerans* JCM 2831	37743	64	38	94	10763	10500	−	
pPRO1	NC_008607	*Pelobacter propionicus* DSM 2379	202397	48	41	94	129679	129957	−	
pRSPA01	NC_009429	*Rhodobacter sphaeroides* ATCC 17025	877879	68	49	97	783519	783791	−	
pSWIT01	NC_009507	*Sphingomonas wittichii* RW1	310228	64	40	95	106554	106820	−	MOB_F
					36	92	35341	35069		
pTcM1	NC_010600	*Acidithiobacillus caldus*	65158	57	56	89	25186	25449	−	MOB_P, MOB_Q
pXCV183	NC_007507	*Xanthomonas campestris* pv. *vesicatoria* str. 85-10	182572	60	33	95	138753	138490	−	
Ti	NC_002377	*Agrobacterium tumefaciens*	194140	55	43	97	180164	180436	−	MOB_Q
Ti plasmid pTiBo542	NC_010929	*Agrobacterium tumefaciens*	244978	55	36	86	209743	209489	−	MOB_Q
					45	98	187204	187479		

[a] This list is the result of a TBLASTN analysis using the amino acid sequence of IHFα or IHFβ as a query under strict conditions (i.e., thresholds of 30% identity and 70% query coverage).

[b] Average G+C content of the plasmid.

[c] Reported TBLASTN identity to IHF.

[d] Plasmid classification according to its source organism (−, Gram-negative plasmid; +, Gram-positive plasmid).

[e] Plasmid classification according to its relaxase gene sequence as described by Garcillán-Barcia et al. [16].

TABLE 4: Plasmids containing the gene encoding Lrp homolog[a].

Plasmid name	Accession no.	Source organism	Length (nt)	G+C content (%)[b]	Identity (%)[c]	Query coverage (%)	Subject start	Subject end	Classification[d]	MOB family[e]
1	NC_008688	*Paracoccus denitrificans* PD1222	653815	67	41	92	252075	251623	–	
					42	93	464218	464673		
					36	96	639341	639811		
					37	85	110140	109724		
A	NC_009007	*Rhodobacter sphaeroides* 2.4.1	114045	69	39	93	30241	29789	–	MOB_F
B	NC_007488	*Rhodobacter sphaeroides* 2.4.1	114178	70	43	96	81861	81385	–	
bglu_1p	NC_012723	*Burkholderia glumae* BGR1	133591	61	36	90	124017	123577	–	
Megaplasmid	NC_008043	*Ruegeria* sp. TM1040	821788	59	41	84	143820	144233	–	
					41	91	687257	687706		
					36	91	734136	733690		
Megaplasmid	NC_007974	*Cupriavidus metallidurans* CH34	2580084	64	44	88	1171245	1170814	–	MOB_V
					40	91	1169702	1169256		
					38	97	1586726	1586250		
Megaplasmid	NC_006569	*Ruegeria pomeroyi* DSS-3	491611	63	36	88	356303	355869	–	MOB_C
Megaplasmid	NC_007336	*Ralstonia eutropha* JMP134	634917	61	35	93	377503	377045	–	
p42e	NC_007765	*Rhizobium etli* CFN 42	505334	62	34	71	255037	255384	–	
p42f	NC_007766	*Rhizobium etli* CFN 42	642517	61	45	88	436907	437341	–	
					43	91	406350	405901		
					41	85	491383	491799		
					39	95	210634	211098		
					39	96	199426	199899		
pAB510a	NC_013855	*Azospirillum* sp. B510	1455109	68	57	88	274908	275342	–	
					44	95	979549	980013		
					32	94	1180335	1179874		
pAB510b	NC_013856	*Azospirillum* sp. B510	723779	67	44	84	471830	472243	–	
					32	94	318139	318600		
pAB510c	NC_013857	*Azospirillum* sp. B510	681723	67	45	85	408064	407645	–	
					34	91	36385	36834		
pAB510d	NC_013858	*Azospirillum* sp. B510	628837	68	44	79	472768	472379	–	
					40	90	323184	322741		
					37	87	281438	281866		
					30	85	619027	618623		
pAtS4e	NC_011981	*Agrobacterium vitis* S4	631775	57	30	87	460443	460871	–	MOB_Q
					34	74	425247	424888		
pBPHY01	NC_010625	*Burkholderia phymatum* STM815	1904893	62	46	85	1153608	1154027	–	
pBPHY02	NC_010627	*Burkholderia phymatum* STM815	595108	59	41	91	271795	271346	–	

Table 4: Continued.

Plasmid name	Accession no.	Source organism	Length (nt)	G+C content (%)[b]	Identity (%)[c]	Query coverage (%)	Subject start	Subject end	Classification[d]	MOB family[e]
pC	NC_010997	*Rhizobium etli* CIAT 652	1091523	61	46	88	617696	618130	–	MOB$_Q$
					42	90	609059	608619		
					39	95	417738	418202		
					42	79	714804	715193		
					39	93	406570	407025		
pCAUL01	NC_010335	*Caulobacter* sp. K31	233649	67	34	89	182479	182042	–	MOB$_Q$
pEST4011	NC_005793	*Achromobacter denitrificans*	76958	62	58	88	41224	40793	–	MOB$_P$
					58	88	34233	33802		
pGMI1000MP	NC_003296	*Ralstonia solanacearum* GMI1000	2094509	67	43	98	1737958	1738437	–	
					46	93	822030	821572		
pHV4	NC_013966	*Haloferax volcanii* DS2	635786	62	33	71	401763	401410	Archaea	
pIJB1	NC_013666	*Burkholderia cepacia*	99448	63	58	88	74907	75338	–	MOB$_P$
pK2044	NC_006625	*Klebsiella pneumoniae* NTUH-K2044	224152	50	33	90	194643	195086	–	
pLVPK	NC_005249	*Klebsiella pneumoniae*	219385	50	33	90	46236	46679	–	
pMLa	NC_002679	*Mesorhizobium loti* MAFF303099	351911	59	32	93	185603	185148	–	
					30	89	207314	206877		
pMLb	NC_002682	*Mesorhizobium loti* MAFF303099	208315	60	37	93	24632	24177	–	
pNGR234a	NC_000914	*Rhizobium* sp. NGR234	536165	58	41	70	197189	196845	–	MOB$_Q$
					30	89	188867	188430		
pNGR234b	NC_012586	*Rhizobium* sp. NGR234	2430033	62	46	90	656547	656107	–	MOB$_Q$
					45	85	667494	667913		
					43	90	1038020	1038463		
					44	85	682796	683215		
					38	96	2400849	2401319		
					44	79	709104	708715		
					41	89	28336	28761		
					33	89	1108900	1109337		
					36	90	703213	702764		
					32	77	1112953	1112582		
pPNAP04	NC_008760	*Polaromonas naphthalenivorans* CJ2	143747	59	35	90	142511	142068	–	
pR132501	NC_012848	*Rhizobium leguminosarum* bv. *trifolii* WSM1325	828924	60	47	88	234905	234471	–	MOB$_Q$
					44	86	386338	386760		

Table 4: Continued.

Plasmid name	Accession no.	Source organism	Length (nt)	G+C content (%)[b]	Identity (%)[c]	Query coverage (%)	Subject start	Subject end	Classification[d]	MOB family[e]
					39	93	645542	645087		
					42	79	147165	146776		
pRALTA	NC_010529	*Cupriavidus taiwanensis*	557200	60	38	91	465839	465393	–	
pRHL1	NC_008269	*Rhodococcus jostii* RHA1	1123075	65	36	91	854207	854656	+	
					33	84	783666	783253		
pRL12	NC_008378	*Rhizobium leguminosarum* bv. viciae 3841	870021	61	46	88	599116	598682	–	MOB_Q
					43	88	658287	658718		
					39	93	45601	45146		
					42	79	450080	449691		
pRL8	NC_008383	*Rhizobium leguminosarum* bv. viciae 3841	147463	59	33	87	70763	70344	–	MOB_Q
pRLG201	NC_011368	*Rhizobium leguminosarum* bv. trifolii WSM2304	1266105	60	45	89	917573	917136	–	MOB_Q
					44	85	41998	42417		
					44	79	473039	472650		
					40	93	1162146	1161691		
					40	93	1150939	1150484		
					32	88	707587	707162		
pRSKD131A	NC_011962	*Rhodobacter sphaeroides* KD131	157345	70	42	96	148295	147819	–	
pRSKD131B	NC_011960	*Rhodobacter sphaeroides* KD131	103355	70	39	93	98400	97948	–	
pRSPA01	NC_009429	*Rhodobacter sphaeroides* ATCC 17025	877879	68	40	90	31309	30866	–	
					39	88	659383	658952		
pRSPH01	NC_009040	*Rhodobacter sphaeroides* ATCC 17029	122606	70	39	93	118088	118540	–	
pSMED01	NC_009620	*Sinorhizobium medicae* WSM419	1570951	61	40	77	143180	143557	–	MOB_Q
					34	89	574284	573847		
pSMED02	NC_009621	*Sinorhizobium medicae* WSM419	1245408	60	42	91	556486	556932	–	MOB_Q
					40	91	842324	842758		
					31	87	22345	21917		
pSMED03	NC_009622	*Sinorhizobium medicae* WSM419	219313	60	46	95	105044	105508	–	

Table 4: Continued.

Plasmid name	Accession no.	Source organism	Length (nt)	G+C content (%)[b]	Identity (%)[c]	Query coverage (%)	Subject start	Subject end	Classification[d]	MOB family[e]
pSmeSM11a	NC_013545	*Sinorhizobium meliloti*	144170	60	46	96	70449	70922	–	MOB_Q
pSymA	NC_003037	*Sinorhizobium meliloti* 1021	1354226	60	43	89	1060699	1060262	–	MOB_Q
pSymB	NC_003078	*Sinorhizobium meliloti* 1021	1683333	62	38	90	440778	440335	–	MOB_Q
					36	89	29555	29992		
pTiS4	NC_011982	*Agrobacterium vitis* S4	258824	57	42	79	96920	97309	–	MOB_Q
Unnamed	NC_011961	*Thermomicrobium roseum* DSM 5159	917738	66	30	85	736739	737146	–	MOB_P

[a] This list is the result of a TBLASTN analysis using the amino acid sequence of Lrp as a query under strict conditions (i.e., thresholds of 30% identity and 70% query coverage).

[b] Average G+C content of the plasmid.

[c] Reported TBLASTN identity to Lrp.

[d] Plasmid classification according to its source organism (−, Gram-negative plasmid; +, Gram-positive plasmid).

[e] Plasmid classification according to its relaxase gene sequence as described by Garcillán-Barcia et al. [16].

TABLE 5: Plasmids containing the gene encoding MvaT or NdpA homolog[a].

Plasmid name	Accession no.	Source organism	Length (nt)	G+C content (%)[b]	Identity (%)[c]	Query coverage (%)	Subject start	Subject end	Classification[d]	MOB family[e]
					MvaT					
pCAR1	NC_004444	*Pseudomonas resinovorans*	199035	56	61	98	77640	77993	–	MOB_H
pQBR103	NC_009444	*Pseudomonas fluorescens* SBW25	425094	53	61	96	98076	97717	–	
pWW53	NC_008275	*Pseudomonas putida*	107929	57	61	98	8415	8768	–	
					NdpA					
p0908	NC_010113	*Vibrio* sp. 0908	81413	49	51	99	79731	78736	–	
pCAR1	NC_004444	*Pseudomonas resinovorans*	199035	56	36	98	95390	94395	–	MOB_H
pQBR103	NC_009444	*Pseudomonas fluorescens* SBW25	425094	53	31	99	161413	160400	–	

[a] This list is the result of a TBLASTN analysis using the amino acid sequence of MvaT or NdpA as a query under strict conditions (i.e., thresholds of 30% identity and 70% query coverage).

[b] Average G+C content of the plasmid.

[c] Reported TBLASTN identity to MvaT or NdpA.

[d] Plasmid classification according to its source organism (−, Gram-negative plasmid).

[e] Plasmid classification according to its relaxase gene sequence as described by Garcillán-Barcia et al. [16].

Table 6: Gram-negative plasmids containing the gene encoding Hha-like protein[a].

Plasmid name	Accession no.	Source organism	Length (nt)	G+C content (%)[b]	NAP gene homolog	Identity (%)[c]	Query coverage (%)	Subject start	Subject end	MOB family[d]
55989p	NC_011752	*Escherichia coli* 55989	72482	46		53	92	10025	9828	
NR1	NC_009133	*Escherichia coli*	94289	52		53	92	87193	87390	MOB_F
p1658/97	NC_004998	*Escherichia coli*	125491	51		55	92	36419	36616	MOB_F
p1ESCUM	NC_011749	*Escherichia coli* UMN026	122301	50		53	92	53508	53311	MOB_F
p2ESCUM	NC_011739	*Escherichia coli* UMN026	33809	42		62	90	7682	7488	MOB_Q
p53638_226	NC_010719	*Escherichia coli* 53638	225683	48		55	92	67615	67418	MOB_F
pAPEC-O1-R	NC_009838	*Escherichia coli* APEC O1	241387	46		50	92	61389	61586	MOB_H
pAPEC-O2-ColV	NC_007675	*Escherichia coli*	184501	49		55	92	3882	3685	MOB_F
pAPEC-O2-R	NC_006671	*Escherichia coli*	101375	53		53	92	4856	4659	MOB_F
pBS512_211	NC_010660	*Shigella boydii* CDC 3083-94	210919	46		55	89	190719	190910	MOB_F
pBS512_33	NC_010657	*Shigella boydii* CDC 3083-94	33103	41		62	90	2894	3088	
pC15-1a	NC_005327	*Escherichia coli*	92353	53		53	92	87490	87687	MOB_F
pCP301	NC_004851	*Shigella flexneri* 2a str. 301	221618	46		55	92	207828	208025	MOB_F
pCROD1	NC_013717	*Citrobacter rodentium* ICC168	54449	47		56	92	53220	53417	
pCROD2	NC_013718	*Citrobacter rodentium* ICC168	39265	42		62	90	15526	15332	
pCT02021853_74	NC_011204	*Salmonella enterica* subsp. *enterica* serovar Dublin str. CT_02021853	74551	49		62	90	48482	48288	MOB_Q
pCTX-M3	NC_004464	*Citrobacter freundii*	89468	51		38	71	26136	26294	MOB_P
			89468			31	96	40648	40439	
pCTXM360	NC_011641	*Klebsiella pneumoniae*	68018	51		38	71	64551	64709	MOB_P
			68018			31	96	10927	10718	
pCVM29188_146	NC_011076	*Salmonella enterica* subsp. *enterica* serovar Kentucky str. CVM29188	146811	49		53	92	18755	18558	MOB_F
pEC14_114	NC_013175	*Escherichia coli*	114222	51		53	92	113985	114182	MOB_F
pEC-IMP	NC_012555	*Enterobacter cloacae*	318782	48	H-NS	50	92	60491	60688	MOB_H
pEC-IMPQ	NC_012556	*Enterobacter cloacae*	324503	48	H-NS	50	92	60491	60688	MOB_H
pEG356	NC_013727	*Shigella sonnei*	70275	52		53	92	69444	69641	MOB_F
pEK499	NC_013122	*Escherichia coli*	117536	53		53	92	41985	42182	
pEK516	NC_013121	*Escherichia coli*	64471	53		53	92	31410	31213	

Table 6: Continued.

Plasmid name	Accession no.	Source organism	Length (nt)	G+C content (%)[b]	NAP gene homolog	Identity (%)[c]	Query coverage (%)	Subject start	Subject end	MOB family[d]
pEL60	NC_005246	*Erwinia amylovora*	60145	51		38	71	23187	23345	MOB_P
			60145			31	96	37863	37654	
pEntH10407	NC_013507	*Escherichia coli ETEC* H10407	67094	51		55	78	43421	43254	MOB_F
pHCM1	NC_003384	*Salmonella enterica* subsp. *enterica* serovar Typhi str. CT18	218160	48	H-NS	47	100	105911	106117	MOB_H
pK2044	NC_006625	*Klebsiella pneumoniae* NTUH-K2044	224152	50	H-NS, Lrp	45	85	143331	143528	
pK29	NC_010870	*Klebsiella pneumoniae*	269674	46		50	92	59322	59519	MOB_H
pKF3-70	NC_013542	*Klebsiella pneumoniae*	70057	52		53	92	15967	15770	MOB_F
pKF3-94	NC_013950	*Klebsiella pneumoniae*	94219	52		58	96	9596	9390	MOB_F
pKP187	NC_011282	*Klebsiella pneumoniae* 342	187922	47		64	96	110083	109877	
			187922			42	89	1550	1344	
pKPN3	NC_009649	*Klebsiella pneumoniae* subsp. *pneumoniae* MGH 78578	175879	52		59	97	56930	56721	MOB_F
plasmid_153 kb	NC_009705	*Yersinia pseudotuberculosis* IP 31758	153140	40	H-NS	69	93	63342	63542	
			153140			56	92	49734	49931	
pLVPK	NC_005249	*Klebsiella pneumoniae*	219385	50	H-NS, Lrp	61	97	148056	147847	
			219385			45	85	214828	215025	
pMAK1	NC_009981	*Salmonella enterica* subsp. *enterica* serovar Choleraesuis	208409	47	H-NS	47	100	49208	49414	MOB_H
pMAS2027	NC_013503	*Escherichia coli*	42644	43		62	90	19685	19491	MOB_Q
pO103	NC_013354	*Escherichia coli* O103:H2 str. 12009	75546	49		55	92	51727	51924	MOB_F
pO111_1	NC_013365	*Escherichia coli* O111:H- str. 11128	204604	47	H-NS	47	100	66925	67131	MOB_H
pO111_3	NC_013366	*Escherichia coli* O111:H- str. 11128	77690	50		55	92	11975	12172	MOB_F
pO157	NC_013010	*Escherichia coli* O157:H7 str. TW14359	94601	48		55	92	70792	70989	
pO157	NC_011350	*Escherichia coli* O157:H7 str. EC4115	94644	48		55	92	54735	54932	
pO157	NC_007414	*Escherichia coli* O157:H7 EDL933	92077	48		55	92	1667	1864	
pO157	NC_002128	*Escherichia coli* O157:H7 str. Sakai	92721	48		55	92	71183	71380	

Table 6: Continued.

Plasmid name	Accession no.	Source organism	Length (nt)	G+C content (%)[b]	NAP gene homolog	Identity (%)[c]	Query coverage (%)	Subject start	Subject end	MOB family[d]
pO26I	NC_011812	*Escherichia coli*	72946	51		53	92	66608	66805	MOB_F
pO86A1	NC_008460	*Escherichia coli*	120730	49		55	92	101598	101795	MOB_F
pOLA52	NC_010378	*Escherichia coli*	51602	46		62	90	12114	11920	MOB_Q
pOU1114	NC_010421	*Salmonella enterica* subsp. *enterica* serovar Dublin	34595	41		62	90	5446	5252	MOB_Q
pOU1115	NC_010422	*Salmonella enterica* subsp. *enterica* serovar Dublin	74589	49		62	90	37246	37052	MOB_Q
pSB4_227	NC_007608	*Shigella boydii* Sb227	126697	47		55	92	110688	110885	MOB_F
pSE11-1	NC_011419	*Escherichia coli* SE11	100021	50		56	92	58407	58210	MOB_P
pSE34	NC_010860	*Salmonella enterica* subsp. *enterica* serovar Enteritidis	32950	41		62	90	21875	22069	MOB_Q
pSFO157	NC_009602	*Escherichia coli*	121239	50		52	75	1709	1870	MOB_F
pSG1	NC_007713	*Sodalis glossinidius* str. "*morsitans*"	83306	49	H-NS	48	92	2294	2491	
pSG1	NC_007182	*Sodalis glossinidius*	81553	49		48	92	56217	56414	
pSMS35_130	NC_010488	*Escherichia coli* SMS-3-5	130440	51		55	92	3364	3167	MOB_F
pSS_046	NC_007385	*Shigella sonnei* Ss046	214396	45		55	92	178363	178560	MOB_F
pUTI89	NC_007941	*Escherichia coli* UTI89	114230	51		53	92	113993	114190	MOB_F
pWR501	NC_002698	*Shigella flexneri*	221851	46		55	92	207534	207731	MOB_F
R100	NC_002134	*Shigella flexneri* 2b str. 222	94281	52		53	92	87185	87382	MOB_F
R27	NC_002305	*Salmonella enterica* subsp. *enterica* serovar Typhi	180461	46	H-NS	47	100	159402	159196	MOB_H
R478	NC_005211	*Serratia marcescens*	274762	46	H-NS	50	92	59426	59623	MOB_H
R721	NC_002525	*Escherichia coli*	75582	43		66	90	35285	35091	
Unnamed	NC_011148	*Salmonella enterica* subsp. *enterica* serovar Agona str. SL483	37978	41	H-NS	42	93	1363	1163	

[a] This list is the result of a TBLASTN analysis using the amino acid sequence of Hha as a query under strict conditions (i.e., thresholds of 30% identity and 70% query coverage).

[b] Average G+C content of the plasmid.

[c] Reported TBLASTN identity to Hha.

[d] Plasmid classification according to its relaxase gene sequence as described by Garcillán-Barcia et al. [16].

Table 7: MOB$_H$-family plasmids of Gram-negative origin[a].

Plasmid name	Accession no.	Source organism	Length (nt)	G+C content (%)[b]	NAP gene homolog	Identity (%)[c]	Query coverage (%)	Subject start	Subject end
1	NC_007949	*Polaromonas* sp. JS666	360405	57	HU	52	99	61052	60786
1	NC_008573	*Shewanella* sp. ANA-3	278942	46					
2	NC_007950	*Polaromonas* sp. JS666	338007	60					
ICEhin1056	NC_011409	*Haemophilus influenzae*	59393	39					
pAM04528	NC_012693	*Salmonella enterica*	158213	52	HU	57	99	14067	14333
pAPEC-O1-R	NC_009838	*Escherichia coli* APEC O1	241387	46					
pAR060302	NC_012692	*Escherichia coli*	166530	53	HU	57	99	15755	16021
pAsa4	NC_009349	*Aeromonas salmonicida* subsp. *salmonicida* A449	166749	53	HU	60	99	26844	26578
pCAR1	NC_004444	*Pseudomonas resinovorans*	199035	56	MvaT	61	98	77640	77993
					NdpA	36	98	95390	94395
					HU	42	99	97809	98075
pEC-IMP	NC_012555	*Enterobacter cloacae*	318782	48	H-NS	64	99	109370	108969
pEC-IMPQ	NC_012556	*Enterobacter cloacae*	324503	48	H-NS	64	99	109370	108969
peH4H	NC_012690	*Escherichia coli*	148105	53	HU	57	99	14067	14333
pHCM1	NC_003384	*Salmonella enterica* subsp. *enterica* serovar Typhi str. CT18	218160	48	H-NS	61	99	131861	131460
pIP1202	NC_009141	*Yersinia pestis* bv. Orientalis str. IP275	182913	53	HU	57	99	14067	14333
pK29	NC_010870	*Klebsiella pneumoniae*	269674	46					
plasmid1	NC_007901	*Rhodoferax ferrireducens* T118	257447	54					
pMAK1	NC_009981	*Salmonella enterica* subsp. *enterica* serovar Choleraesuis	208409	47	H-NS	61	99	60046	59645
pMAQU02	NC_008739	*Marinobacter aquaeolei* VT8	213290	53					
pO111_1	NC_013365	*Escherichia coli* O111:H- str. 11128	204604	47	H-NS	61	99	80175	79774
pP91278	NC_008613	*Photobacterium damselae* subsp. *Piscicida*	131520	52	HU	57	99	125918	126184

Table 7: Continued.

Plasmid name	Accession no.	Source organism	Length (nt)	G+C content (%)[b]	NAP gene homolog	Identity (%)[c]	Query coverage (%)	Subject start	Subject end
pP99-018	NC_008612	*Photobacterium damselae* subsp. *piscicida*	150157	51	HU	57	99	133314	133580
pRA1	NC_012885	*Aeromonas hydrophila*	143963	51	HU	58	99	15573	15839
pRp12D01	NC_012855	*Ralstonia pickettii* 12D	389779	58	HU	37	99	321346	321080
pSN254	NC_009140	*Salmonella enterica* subsp. *enterica* serovar Newport str. SL254	176473	53	HU	57	99	14067	14333
pTK9001	NC_013930	*Thioalkalivibrio* sp. K90mix	240256	62					
pYR1	NC_009139	*Yersinia ruckeri*	158038	51	HU	57	99	15070	15336
R27	NC_002305	*Salmonella enterica* subsp. *enterica* serovar Typhi	180461	46	H-NS	61	99	148225	148626
R478	NC_005211	*Serratia marcescens*	274762	46	H-NS	64	99	111747	111346
Rts1	NC_003905	*Proteus vulgaris*	217182	46					

[a] This list is the result of a TBLASTN analysis using the 300 N-terminal amino acid sequence of protein TraI_R27 as a query under strict conditions (i.e., thresholds of 30% identity and 70% query coverage).

[b] Average G+C content of the plasmid.

[c] Reported TBLASTN identity to each NAP.

Table 8: MOB$_Q$-family plasmids of Gram-negative origin[a].

Plasmid name	Accession no.	Source organism	Length (nt)	G+C content (%)[b]	NAP gene homolog	Identity (%)[c]	Query coverage (%)	Subject start	Subject end
1	NC_008242	*Chelativorans* sp. BNC1	343931	62	HU	41	94	133932	133678
3	NC_007617	*Nitrosospira multiformis* ATCC 25196	14159	50					
3	NC_007961	*Nitrobacter hamburgensis* X14	121408	62					
At	NC_003064	*Agrobacterium tumefaciens* str. C58	542868	57	IHF	36	82	112654	112412
C	NC_010542	*Cyanothece* sp. ATCC 51142	14685	38					
ColE9-J	NC_011977	*Escherichia coli*	7577	50					
DN1	NC_002636	*Dichelobacter nodosus*	5112	62					
F plasmid	NC_008036	*Sphingopyxis alaskensis* RB2256	28543	60					
p11745	NC_013546	*Actinobacillus pleuropneumoniae*	5486	38					
p12494	NC_010889	*Actinobacillus pleuropneumoniae*	14393	33					
p1ABAYE	NC_010401	*Acinetobacter baumannii* AYE	5644	35					
p1META1	NC_012807	*Methylobacterium extorquens* AM1	44195	68					
p1METDI	NC_012987	*Methylobacterium extorquens* DM4	141504	65					
p2007057	NC_011897	*Salmonella enterica* subsp. *enterica* serovar Bovismorbificans	4270	47					
p2ABSDF	NC_010396	*Acinetobacter baumannii* SDF	25014	35					
p2ESCUM	NC_011739	*Escherichia coli* UMN026	33809	42					
p2META1	NC_012809	*Methylobacterium extorquens* AM1	37858	65	IHF	44	95	28369	28635
p3ABSDF	NC_010398	*Acinetobacter baumannii* SDF	24922	34					
p42a	NC_007762	*Rhizobium etli* CFN 42	194229	58					
p49879.1	NC_006907	*Leptospirillum ferrooxidans*	28878	58	HU	47	99	3281	3015

Table 8: Continued.

Plasmid name	Accession no.	Source organism	Length (nt)	G+C content (%)^b	NAP gene homolog	Identity (%)^c	Query coverage (%)	Subject start	Subject end
p49879.2	NC_006909	*Leptospirillum ferrooxidans*	28012	55	HU	48	99	15858	15592
pAb5S9	NC_009476	*Aeromonas bestiarum*	24716	54					
pACRY07	NC_009473	*Acidiphilium cryptum* JF-5	5629	58					
pAgK84	NC_011994	*Agrobacterium radiobacter* K84	44420	54					
pAM5	NC_008691	*Acidiphilium multivorum*	5161	58					
pAMI2	NC_010847	*Paracoccus aminophilus*	18563	62					
pAMI3	NC_013513	*Paracoccus aminophilus*	5575	61					
pAPA01-030	NC_013212	*Acetobacter pasteurianus* IFO 3283-01	49961	54					
pAPA01-040	NC_013213	*Acetobacter pasteurianus* IFO 3283-01	3204	54					
pAtK84b	NC_011990	*Agrobacterium radiobacter* K84	184668	59	IHF	38	86	54109	53855
pAtS4b	NC_011991	*Agrobacterium vitis* S4	130435	56	IHF	47	97	44880	45152
pAtS4c	NC_011984	*Agrobacterium vitis* S4	211620	59	HU	45	94	141245	140991
pAtS4e	NC_011981	*Agrobacterium vitis* S4	631775	57	HU	41	94	40476	40222
					Lrp	30	87	460443	460871
					Lrp	34	74	425247	424888
pAV2	NC_010310	*Acinetobacter venetianus*	15135	36					
pB	NC_010996	*Rhizobium etli* CIAT 652	429111	58					
pBGR3	NC_012847	*Bartonella grahamii* as4aup	28192	36					
pBS512_5	NC_010659	*Shigella boydii* CDC 3083-94	5114	46					
pC	NC_010997	*Rhizobium etli* CIAT 652	1091523	61	Lrp	46	88	617696	618130
					Lrp	42	90	609059	608619
					Lrp	39	95	417738	418202
					Lrp	42	79	714804	715193
					Lrp	39	93	406570	407025
pCAUL01	NC_010335	*Caulobacter* sp. K31	233649	67	HU	44	99	97598	97329
					Lrp	34	89	182479	182042

Table 8: Continued.

Plasmid name	Accession no.	Source organism	Length (nt)	G+C content (%)[b]	NAP gene homolog	Identity (%)[c]	Query coverage (%)	Subject start	Subject end
pCAUL02	NC_010333	*Caulobacter* sp. K31	177878	64					
pCCK1900	NC_011378	*Pasteurella multocida*	10226	61					
pCCK381	NC_006994	*Pasteurella multocida*	10874	61					
pCFPG4	NC_011563	*Candidatus Azobacteroides pseudotrichonympha* genomovar. CFP2	4149	44					
pCHE-A	NC_012006	*Enterobacter cloacae*	7560	60					
pColE8	NC_012882	*Escherichia coli*	6751	51					
pCROD3	NC_013719	*Citrobacter rodentium* ICC168	3910	51					
pCT02021853_74	NC_011204	*Salmonella enterica* subsp. *enterica* serovar Dublin str. CT_02021853	74551	49					
pCVM19633_4	NC_011093	*Salmonella enterica* subsp. *enterica* serovar Schwarzengrund str. CVM19633	4585	48					
pDSHI01	NC_009955	*Dinoroseobacter shibae* DFL 12	190506	60					
pET09	NC_010695	*Erwinia tasmaniensis* Et1/99	9299	47					
pGDIA01	NC_011367	*Gluconacetobacter diazotrophicus* PAl 5	27455	59					
pGOX3	NC_006674	*Gluconobacter oxydans* 621H	14547	56					
pHCG3	NC_005873	*Oligotropha carboxidovorans* OM5	133058	61					
pHRM2a	NC_012109	*Desulfobacterium autotrophicum* HRM2	68709	42					
pIGJC156	NC_009781	*Escherichia coli*	5146	47					
pIGMS5	NC_010883	*Escherichia coli*	6750	51					
pIGWZ12	NC_010885	*Escherichia coli*	4072	50					
pISP3	NC_013970	*Sphingomonas* sp. MM-1	43398	63					
pJD4	NC_002098	*Neisseria gonorrhoeae*	7426	38					
plasmid1	NC_007801	*Jannaschia* sp. CCS1	86072	58					

Table 8: Continued.

Plasmid name	Accession no.	Source organism	Length (nt)	G+C content (%)[b]	NAP gene homolog	Identity (%)[c]	Query coverage (%)	Subject start	Subject end
pLD-TEX-KL	NC_009966	*Fluoribacter dumoffii*	66512	39					
pMAC	NC_006877	*Acinetobacter baumannii*	9540	35					
pMAS2027	NC_013503	*Escherichia coli*	42644	43					
pMbo4.6	NC_013500	*Moraxella bovis*	4658	39					
pMCHL01	NC_011758	*Methylobacterium chloromethanicum* CM4	380207	66					
pMG160	NC_004527	*Rhodobacter blasticus*	3431	67					
pMG828-2	NC_008487	*Escherichia coli*	4091	50					
pMG828-4	NC_008489	*Escherichia coli*	7462	48					
pMMCU1	NC_013056	*Acinetobacter calcoaceticus*	8771	35					
pMMCU2	NC_013506	*Acinetobacter baumannii*	10270	36					
pMRAD01	NC_010510	*Methylobacterium radiotolerans* JCM 2831	586164	70					
pMS260	NC_005312	*Actinobacillus pleuropneumoniae*	8124	61					
pNGR234a	NC_000914	*Rhizobium* sp. NGR234	536165	59	Lrp	41	70	197189	196845
					Lrp	30	89	188867	188430
pNGR234b	NC_012586	*Rhizobium* sp. NGR234	2430033	62	Lrp	46	90	656547	656107
					Lrp	45	85	667494	667913
					Lrp	43	90	1038020	1038463
					Lrp	44	85	682796	683215
					Lrp	38	96	2400849	2401319
					Lrp	44	79	709104	708715
					Lrp	41	89	28336	28761
					Lrp	33	89	1108900	1109337
					Lrp	36	90	703213	702764
					Lrp	32	77	1112953	1112582
pNL2	NC_009427	*Novosphingobium aromaticivorans* DSM 12444	487268	66					
pO111_4	NC_013367	*Escherichia coli* O111:H-str. 11128	8140	50					
pO26-S4	NC_011228	*Escherichia coli*	6758	51					
pOLA52	NC_010378	*Escherichia coli*	51602	46					
pOU1114	NC_010421	*Salmonella enterica* subsp. *enterica* serovar Dublin	34595	42					

Table 8: Continued.

Plasmid name	Accession no.	Source organism	Length (nt)	G+C content (%)[b]	NAP gene homolog	Identity (%)[c]	Query coverage (%)	Subject start	Subject end
pOU1115	NC_010422	*Salmonella enterica* subsp. *enterica* serovar Dublin	74589	49					
pP	NC_003455	*Salmonella enterica* subsp. *enterica* serovar *Enteritidis*	4301	50					
pP742405	NC_011733	*Cyanothece* sp. PCC 7424	18083	38					
pP742406	NC_011734	*Cyanothece* sp. PCC 7424	15219	40					
pPMA4326C	NC_005921	*Pseudomonas syringae* pv. *maculicola*	8244	53					
pPNAP07	NC_008763	*Polaromonas naphthalenivorans* CJ2	9898	57					
pPRO2	NC_008608	*Pelobacter propionicus* DSM 2379	30722	56					
pPT1	NC_002143	*Comamonas testosteroni*	15398	56					
pR132501	NC_012848	*Rhizobium leguminosarum* bv. trifolii WSM1325	828924	60	Lrp	47	88	234905	234471
					Lrp	44	86	386338	386760
					Lrp	39	93	645542	645087
					Lrp	42	79	147165	146776
pR132502	NC_012858	*Rhizobium leguminosarum* bv. trifolii WSM1325	660973	61					
pR132503	NC_012853	*Rhizobium leguminosarum* bv. trifolii WSM1325	516088	59	HU	47	94	300662	300916
pR132504	NC_012852	*Rhizobium leguminosarum* bv. trifolii WSM1325	350312	61					
pR132505	NC_012854	*Rhizobium leguminosarum* bv. trifolii WSM1325	294782	60					
pRF	NC_007110	*Rickettsia felis* URRWXCal2	62829	34					
pRFdelta	NC_007111	*Rickettsia felis* URRWXCal2	39263	33					
pRi1724	NC_002575	*Agrobacterium rhizogenes*	217594	57					
pRi2659	NC_010841	*Agrobacterium rhizogenes*	185462	58					
pRL10	NC_008381	*Rhizobium leguminosarum* bv. viciae 3841	488135	60					
pRL11	NC_008384	*Rhizobium leguminosarum* bv. viciae 3841	684202	61					
pRL12	NC_008378	*Rhizobium leguminosarum* bv. viciae 3841	870021	61	Lrp	46	88	599116	598682

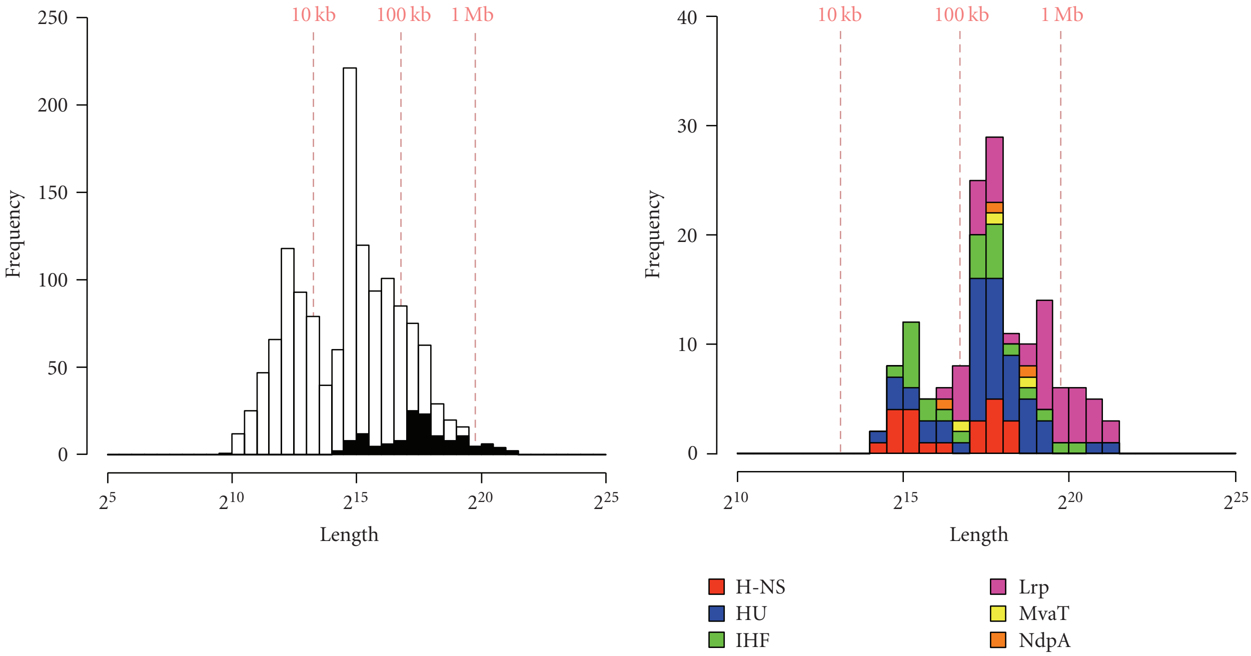
Table 8: Continued.

Plasmid name	Accession no.	Source organism	Length (nt)	G+C content (%)[b]	NAP gene homolog	Identity (%)[c]	Query coverage (%)	Subject start	Subject end
					Lrp	43	88	658287	658718
					Lrp	39	93	45601	45146
					Lrp	42	79	450080	449691
pRL7	NC_008382	*Rhizobium leguminosarum* bv. viciae 3841	151564	58	HU	48	94	20484	20230
pRL8	NC_008383	*Rhizobium leguminosarum* bv. viciae 3841	147463	59	Lrp	33	87	70763	70344
pRLG201	NC_011368	*Rhizobium leguminosarum* bv. trifolii WSM2304	1266105	60	Lrp	45	89	917573	917136
					Lrp	44	85	41998	42417
					Lrp	44	79	473039	472650
					Lrp	40	93	1162146	1161691
					Lrp	40	93	1150939	1150484
					Lrp	32	88	707587	707162
pRM	NC_010927	*Rickettsia monacensis*	23486	32					
pSC101	NC_002056	*Salmonella enterica* subsp. *enterica* serovar Typhimurium	9263	51					
pSE11-6	NC_011411	*Escherichia coli* SE11	4082	49					
pSE34	NC_010860	*Salmonella enterica* subsp. *enterica* serovar Enteritidis	32950	41					
pSMED01	NC_009620	*Sinorhizobium medicae* WSM419	1570951	62	Lrp	40	77	143180	143557
					Lrp	34	89	574284	573847
pSMED02	NC_009621	*Sinorhizobium medicae* WSM419	1245408	60	Lrp	42	91	556486	556932
					Lrp	40	91	842324	842758
					Lrp	31	87	22345	21917
pSmeSM11a	NC_013545	*Sinorhizobium meliloti*	144170	60	Lrp	46	96	70449	70922
pSmeSM11b	NC_010865	*Sinorhizobium meliloti* SM11	181251	59					
pSMS35_4	NC_010486	*Escherichia coli* SMS-3-5	4074	50					
pSx-Qyy	NC_006826	*Sphingobium xenophagum*	5683	56					
pSymA	NC_003037	*Sinorhizobium meliloti* 1021	1354226	60	Lrp				
pSymB	NC_003078	*Sinorhizobium meliloti* 1021	1683333	62	Lrp	38	90	440778	440335

Table 8: Continued.

Plasmid name	Accession no.	Source organism	Length (nt)	G+C content (%)[b]	NAP gene homolog	Identity (%)[c]	Query coverage (%)	Subject start	Subject end
					Lrp	36	89	29555	29992
pTB3	NC_008388	*Roseobacter denitrificans* OCh 114	16575	55					
pTcM1	NC_010600	*Acidithiobacillus caldus*	65158	57	IHF	56	89	25186	25449
pTiS4	NC_011982	*Agrobacterium vitis* S4	258824	57	HU	41	94	27356	27102
					HU	40	94	83408	83154
					Lrp	42	79	96920	97309
pTi-SAKURA	NC_002147	*Agrobacterium tumefaciens*	206479	56	HU	44	94	95763	95509
pUT1	NC_014005	*Sphingobium japonicum* UT26S	31776	64					
pUT2	NC_014009	*Sphingobium japonicum* UT26S	5398	61					
pXAUT01	NC_009717	*Xanthobacter autotrophicus* Py2	316164	65					
pXCV19	NC_007505	*Xanthomonas campestris* pv. *vesicatoria* str. 85-10	19146	60					
pXF51	NC_002490	*Xylella fastidiosa* 9a5c	51158	50					
pYAN-1	NC_008246	*Sphingobium yanoikuyae*	5182	62					
pYAN-2	NC_008247	*Sphingobium yanoikuyae*	4924	64					
RSF1010	NC_001740	*Escherichia coli*	8684	61					
Symbiotic plasmid p42d	NC_004041	*Rhizobium etli* CFN 42	371254	58					
Ti	NC_002377	*Agrobacterium tumefaciens*	194140	55	IHF	43	97	180164	180436
Ti	NC_003065	*Agrobacterium tumefaciens* str. C58	214233	57	HU	44	94	139735	139481
Ti plasmid pTiBo542	NC_010929	*Agrobacterium tumefaciens*	244978	55	IHF	36	86	209743	209489
					IHF	45	98	187204	187479
Unnamed	NC_011143	*Phenylobacterium zucineum* HLK1	382976	69					

[a] This list is the result of a TBLASTN analysis using the 300 N-terminal amino acid sequence of protein MobA_RSF1010 as a query under strict conditions (i.e., thresholds of 30% identity and 70% query coverage).
[b] Average G+C content of the plasmid.
[c] Reported TBLASTN identity to each NAP.



(a)

(b)

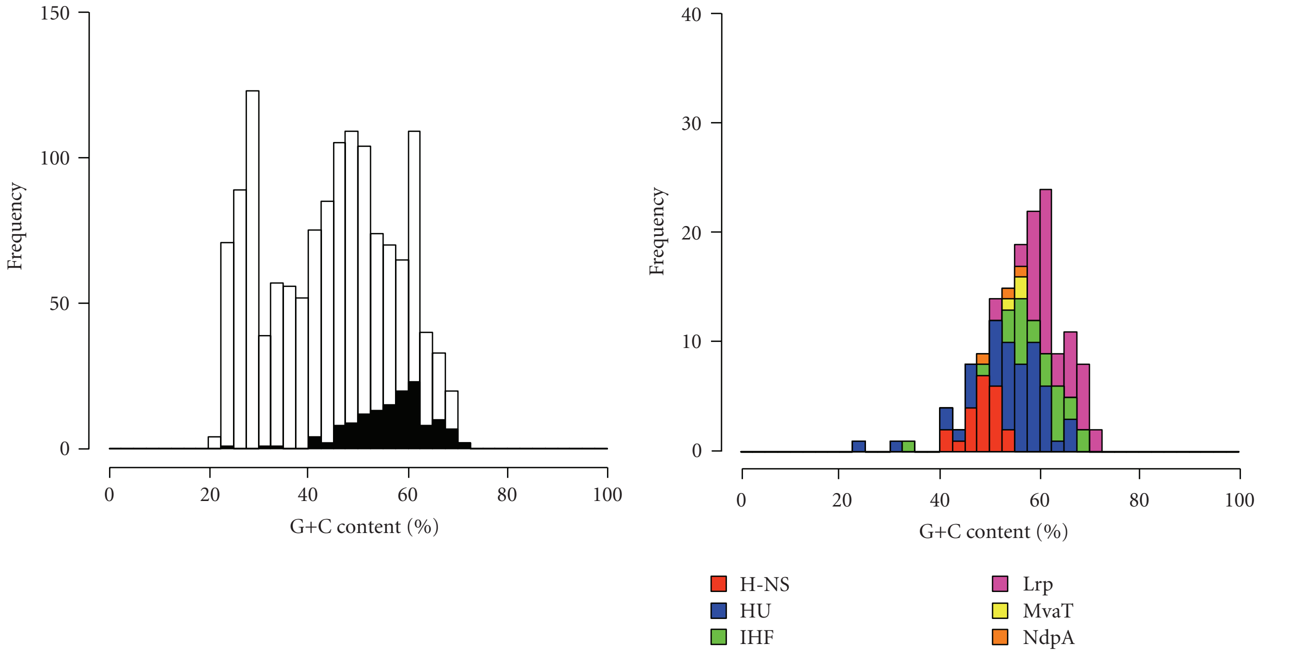
Figure 1: Size comparison of the Gram-negative plasmids with and without NAP gene homologs. (a) A total of 136 Gram-negative plasmids with one or more NAP gene homologs and 1246 Gram-negative plasmids without NAP gene homologs are shown by black and white bars, respectively. (b) Gram-negative plasmids with each NAP gene homolog are as follows: H-NS, red; HU, blue; IHF, green; Lrp, purple; MvaT, yellow; and NdpA, orange.

(83 kb). These results suggest that larger plasmids, especially >100 kb, frequently have NAP gene homologs. Carrying large plasmids may reduce host fitness more than carrying small plasmids because the former have more genes that can disrupt transcriptional networks in the host cell. In addition, large plasmids may have more binding sites for NAPs than small plasmids. Because chromosome-encoded NAPs bind to both chromosomes and plasmids, carrying large plasmids may also result in a reduction in the binding of NAPs to the host chromosome, causing undesirable effects on the host cell. Plasmid-encoded NAP homologs may interact with chromosome-encoded NAPs, coordinately sustain the structure of both chromosome and plasmid, and regulate the transcriptional regulation network [23]. In fact, recent studies have shown that some plasmid-encoded NAP homologs can complement the depletion of chromosomal NAPs and optimize gene transcription both on plasmids and in the host chromosome [14, 15, 24]. Thus, larger plasmids may have NAP gene homologs to maintain host cell fitness. In addition, the average size of the 38 plasmids containing more than one NAP gene homolog was larger (790 kb) than that of the 98 plasmids containing only one NAP gene homolog (199 kb). This suggests that particularly large plasmids have many NAP gene homologs to maintain themselves in the host cell.

Distributions of the NAP genes on proteobacterial genomes were also surveyed using the TBLASTN program. The average size of the completely sequenced bacterial genomes was 3.25 Mb and 1054 NAP genes (100, Fis; 125, H-NS; 236, HU; 247, IHF; 127, Lrp; 119, MvaT; and 100, NdpA) were found in 588 proteobacterial genomes. Frequency of NAP genes in plasmids was higher (1 per 236 kb) than that in proteobacterial genomes (1 per 1.8 Mb), also suggesting that larger plasmids frequently have NAP gene homologs to minimize their negative effects on the host cell.

Of the plasmids with the NAP gene homolog, the average size of those with the H-NS gene homolog was relatively small (132 kb) while that of those with the Lrp gene homolog was relatively large (725 kb). The average sizes of those with the other NAP gene homologs were as follows: HU (301 kb), IHF (230 kb), MvaT (244 kb), and NdpA (235 kb) (Figure 1(b)). H-NS exists in an oligomeric form and binds to DNA, especially A+T-rich regions, by bridging it [25]. This function may be important for regulating gene expression on relatively small plasmids among those with the NAP gene homolog. The activity of H-NS can also be modulated by Hha-like proteins [26]. Intriguingly, TBLASTN analysis showed that 12 (55%) of 22 plasmids with the H-NS gene homolog also carried gene encoding Hha-like protein although only 65 (5%) of all 1382 plasmids carried Hha-like protein gene (Table 6). This suggests the close relationship of H-NS and Hha-like protein. On the other hand, Lrp exists in dimeric, octameric, and hexadecameric forms and compacts DNA by wrapping it [27]. This distinctive DNA-binding ability may be essential for maintaining the structure of particularly larger plasmids.

3.4. Relationships between Plasmid G+C Content and NAP Gene Homolog Distributions. Next, we surveyed the G+C content of the Gram-negative group plasmids with and



(a)

(b)

Figure 2: G+C content comparison of the Gram-negative plasmids with and without NAP gene homologs. (a) A total of 136 Gram-negative plasmids with one or more NAP gene homologs and 1246 Gram-negative plasmids without NAP gene homologs are shown by black and white bars, respectively. (b) Gram-negative plasmids with each NAP gene homolog are as follows: H-NS, red; HU, blue; IHF, green; Lrp, purple; MvaT, yellow; and NdpA, orange.

without NAP gene homologs. The average G+C content of the 136 plasmids with NAP gene homologs was higher (56.4%) than that of all 1382 plasmids (44.8%) (Figure 2(a)). Note that the average G+C content of large and mega plasmids (55.0% and 62.9%, resp.) was higher than that of small and intermediate plasmids (44.8% and 40.4%). Considering that larger plasmids frequently had NAP gene homologs, this seems reasonable. Nevertheless, plasmids with H-NS gene homologs had a lower G+C content (48.3%) than did those with other NAP gene homologs, including HU (54.2%), IHF (58.7%), Lrp (62.3%), MvaT (55.6%), and NdpA (52.9%) (Figure 2(b)). H-NS family protein binds A+T-rich regions not only on chromosomes but also on plasmids [15]. Acquisition of a large A+T-rich plasmid with many H-NS binding sites may result in a reduction in the binding of H-NS to the host chromosome and host cell fitness [14]. It is therefore possible that large A+T-rich plasmids may have to supply another H-NS encoded on themselves to minimize the effect on the host cell. On the other hand, although MvaT-family proteins are the functional homolog of H-NS [10, 15], plasmids containing the MvaT gene homolog were not particularly low in G+C content. Although only three plasmids contained the MvaT gene homolog and thus we cannot discuss this interesting phenomenon in detail, the difference between H-NS and MvaT may be derived from their different origin or host bacteria.

3.5. Relationships between Plasmid Transferability and NAP Gene Homolog Distributions. Conjugative transfer is an essential function of plasmids, through which they play an important role in bacterial evolution and host cell behavior [11, 12]. Relaxase is an essential protein for plasmid transmission involved in the cleavage of the transferring DNA at the origin of transfer (*oriT*) site, and plasmids with relaxase genes are thought to be transmissible. Garcillán-Barcia et al. [16] proposed that transmissible plasmids can be classified into 6 MOB families (MOB_C, MOB_F, MOB_H, MOB_P, MOB_Q, and MOB_V) according to the amino acid sequences of 6 prototype relaxase proteins. MOB_F and MOB_H families are predominantly composed of conjugative plasmids, also called self-transmissible plasmids, and the other 4 families are composed of both mobilizable and conjugative plasmids. Recent studies have reported that plasmid-encoded H-NS family proteins have a “stealth” function and aide horizontal transfer of plasmids [14, 15]. Other NAPs also act as global transcriptional regulators and may regulate expression of genes involved in plasmid transmission. To discuss the relationship between NAP gene homolog distribution and plasmid transferability, we determined the distribution of genes encoding relaxase proteins in Gram-negative plasmids according to the classification by Garcillán-Barcia et al. [16]. Four hundred and nine (30%) of 1382 Gram-negative plasmids carried relaxase genes, and 71 (17%) of those 409 plasmids carried NAP gene homologs. Note that 71 (52%) of 136 plasmids with NAP gene homologs carried relaxase genes. This indicates that plasmids with NAP gene homologs frequently carried the relaxase genes than did those without NAP gene homologs. This phenomenon may be related to the average size of the plasmids. That of the 409 plasmids with relaxase genes was relatively larger (145 kb) than that of all 1382 plasmids (83 kb), corresponding to the fact that larger plasmids frequently had NAP gene homologs.

Four hundred and nine plasmids were classified into each MOB family (13, $\mathrm{MOB_C}$; 128, $\mathrm{MOB_F}$; 29, $\mathrm{MOB_H}$; 86, $\mathrm{MOB_P}$; 131, $\mathrm{MOB_Q}$; and 26, $\mathrm{MOB_V}$). Plasmid 1 (NC_008545) was classified into both the $\mathrm{MOB_C}$ and $\mathrm{MOB_F}$ families. In addition, the $\mathrm{MOB_P}$, $\mathrm{MOB_Q}$, and $\mathrm{MOB_V}$ families were partially overlapped as described by Garcillán-Barcia et al. [16]. Seventy-one plasmids with NAP gene homologs were contained in each MOB family (1, $\mathrm{MOB_C}$; 11, $\mathrm{MOB_F}$; 20, $\mathrm{MOB_H}$; 8, $\mathrm{MOB_P}$; 30, $\mathrm{MOB_Q}$; and 2, $\mathrm{MOB_V}$). Intriguingly, 20 (69%) of 29 $\mathrm{MOB_H}$-family plasmids encoded some NAP homologs, and most of them were H-NS or HU (Table 7). The $\mathrm{MOB_H}$ family was composed of predominantly large conjugative plasmids, such as the IncHI1 group of plasmids, suggesting that HU may also contribute to plasmid transmission as does H-NS. Furthermore, 30 (23%) of 131 $\mathrm{MOB_Q}$-family plasmids also contained some NAP gene homologs, and 15 (50%) of those carried Lrp gene homologs (Table 8). The $\mathrm{MOB_Q}$ family was composed of both mobilizable and conjugative plasmids, such as those of *Rhizobium* and *Agrobacterium*, implying that Lrp may also affect plasmid conjugation. In the other MOB families, plasmids containing NAP gene homologs were less than 10% (8%, $\mathrm{MOB_C}$; 9%, $\mathrm{MOB_F}$; 9%, $\mathrm{MOB_P}$; and 8%, $\mathrm{MOB_V}$). This phenomenon may also be related to the average size of the plasmids contained in each MOB family. $\mathrm{MOB_H}$ (220 kb) and $\mathrm{MOB_Q}$ (198 kb) were larger than $\mathrm{MOB_C}$ (78 kb), $\mathrm{MOB_F}$ (117 kb), $\mathrm{MOB_P}$ (87 kb), and $\mathrm{MOB_V}$ (149 kb). On the other hand, the average G+C content of all plasmids belonging to each MOB family was as follows: $\mathrm{MOB_C}$ (52%), $\mathrm{MOB_F}$ (52%), $\mathrm{MOB_H}$ (51%), $\mathrm{MOB_P}$ (53%), $\mathrm{MOB_Q}$ (54%), and $\mathrm{MOB_V}$ (46%). No relationship between the distribution of NAP gene homologs of each MOB family and the G+C content of plasmids was found.

3.6. Conclusions. We compared the distribution of NAP gene homologs among plasmids and plasmid features. Larger plasmids frequently had NAP gene homologs, possibly to maintain themselves and host cell fitness. Plasmids with NAP gene homologs also frequently carried relaxase genes. Although this may be related to their relatively larger sizes, together with the fact that NAPs affect global gene regulation, it is likely that NAPs contribute to plasmid transmission. Considering the fact that NAPs encoded on plasmids actually help the host cell to integrate newly acquired genes into host regulatory networks [14, 15], large plasmids with NAP gene homologs may be generally more beneficial not only for the host cell, but also for their own existence.

NAP homologs encoded on plasmids can interact with different types of NAPs encoded on the host chromosome and cooperatively regulate host transcriptional networks. Understanding these mechanisms in more detail will shed light on the meanings of the distributions of NAPs on plasmids and chromosomes. Comprehensive analysis of their binding sites in the host and plasmid genomes will help us to understand the relationships between G+C content and the presence of NAPs. Such information will explain how bacteria adapt and evolve by acquiring foreign genes by HGT.

References

[1] C. J. Dorman, "Chapter 2 nucleoid-associated proteins and bacterial physiology," *Advances in Applied Microbiology*, vol. 67, pp. 47–64, 2009.

[2] S. C. Dillon and C. J. Dorman, "Bacterial nucleoid-associated proteins, nucleoid structure and gene expression," *Nature Reviews Microbiology*, vol. 8, no. 3, pp. 185–195, 2010.

[3] M. D. Bradley, M. B. Beach, A. P. J. de Koning, T. S. Pratt, and R. Osuna, "Effects of Fis on *Escherichia coli* gene expression during different growth stages," *Microbiology*, vol. 153, no. 9, pp. 2922–2940, 2007.

[4] W. W. Navarre, S. Porwollik, Y. Wang et al., "Selective silencing of foreign DNA with low GC content by the H-NS protein in *Salmonella*," *Science*, vol. 313, no. 5784, pp. 236–238, 2006.

[5] J. Oberto, S. Nabti, V. Jooste, H. Mignot, and J. Rouviere-Yaniv, "The HU regulon is composed of genes responding to anaerobiosis, acid stress, high osmolarity and SOS induction," *PLoS One*, vol. 4, no. 2, article e4367, 2009.

[6] M. W. Mangan, S. Lucchini, V. Danino, T. Ó. Cróinín, J. C. D. Hinton, and C. J. Dorman, "The integration host factor (IHF) integrates stationary-phase and virulence gene expression in *Salmonella enterica* serovar Typhimurium," *Molecular Microbiology*, vol. 59, no. 6, pp. 1831–1847, 2006.

[7] K. K. Swinger and P. A. Rice, "IHF and HU: flexible architects of bent DNA," *Current Opinion in Structural Biology*, vol. 14, no. 1, pp. 28–35, 2004.

[8] B. K. Cho, C. L. Barrett, E. M. Knight, Y. S. Park, and B. Ø. Palsson, "Genome-scale reconstruction of the Lrp regulatory network in *Escherichia coli*," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, no. 49, pp. 19462–19467, 2008.

[9] L. D. Murphy, J. L. Rosner, S. B. Zimmerman, and D. Esposito, "Identification of two new proteins in spermidine nucleoids isolated from *Escherichia coli*," *Journal of Bacteriology*, vol. 181, no. 12, pp. 3842–3844, 1999.

[10] C. Tendeng, O. A. Soutourina, A. Danchin, and P. N. Bertin, "MvaT proteins in *Pseudomonas* spp.: a novel class of H-NS-like proteins," *Microbiology*, vol. 149, no. 11, pp. 3047–3050, 2003.

[11] L. S. Frost, R. Leplae, A. O. Summers, and A. Toussaint, "Mobile genetic elements: the agents of open source evolution," *Nature Reviews Microbiology*, vol. 3, no. 9, pp. 722–732, 2005.

[12] C. M. Thomas and K. M. Nielsen, "Mechanisms of, and barriers to, horizontal gene transfer between bacteria," *Nature Reviews Microbiology*, vol. 3, no. 9, pp. 711–721, 2005.

[13] A. Carattoli, "Plasmid-mediated antimicrobial resistance in *Salmonella enterica*," *Current Issues in Molecular Biology*, vol. 5, no. 4, pp. 113–122, 2003.

[14] M. Doyle, M. Fookes, AL. Ivens, M. W. Mangan, J. Wain, and C. J. Dorman, "An H-NS-like stealth protein aids horizontal DNA transmission in bacteria," *Science*, vol. 315, no. 5809, pp. 251–252, 2007.

[15] C.-S. Yun, C. Suzuki, K. Naito et al., "Pmr, a histone-like protein H1 (H-NS) family protein encoded by the IncP-7 plasmid pCAR1, is a key global regulator that alters host function," *Journal of Bacteriology*, vol. 192, no. 18, pp. 4720–4731, 2010.

[16] M. P. Garcillán-Barcia, M. V. Francia, and F. de la Cruz, "The diversity of conjugative relaxases and its application in plasmid classification," *FEMS Microbiology Reviews*, vol. 33, no. 3, pp. 657–687, 2009.

[17] J. Wei, M. B. Goldberg, V. Burland et al., "Complete genome sequence and comparative genomics of *Shigella flexneri* serotype 2a strain 2457T," *Infection and Immunity*, vol. 71, no. 5, pp. 2775–2786, 2003.

[18] C. K. Sherburne, T. D. Lawley, M. W. Gilmour et al., "The complete DNA sequence and analysis of R27, a large IncHI plasmid from *Salmonella typhi* that is temperature sensitive for transfer," *Nucleic Acids Research*, vol. 28, no. 10, pp. 2177–2186, 2000.

[19] J. Wain, L. T. D. Nga, C. Kidgell et al., "Molecular analysis of incHI1 antimicrobial resistance plasmids from *Salmonella* serovar Typhi strains associated with typhoid fever," *Antimicrobial Agents and Chemotherapy*, vol. 47, no. 9, pp. 2732–2739, 2003.

[20] A. Tett, A. J. Spiers, L. C. Crossman et al., "Sequence-based analysis of pQBR103; a representative of a unique, transfer-proficient mega plasmid resident in the microbial community of sugar beet," *ISME Journal*, vol. 1, no. 4, pp. 331–340, 2007.

[21] K. Maeda, H. Nojiri, M. Shintani, T. Yoshida, H. Habe, and T. Omori, "Complete nucleotide sequence of carbazole/dioxin-degrading plasmid pCAR1 in *Pseudomonas resinovorans* strain CA10 indicates its mosaicity and the presence of large catabolic transposon Tn*4676*," *Journal of Molecular Biology*, vol. 326, no. 1, pp. 21–33, 2003.

[22] Y. Takahashi, M. Shintani, H. Yamane, and H. Nojiri, "The complete nucleotide sequence of pCAR2: pCAR2 and pCAR1 were structurally identical incP-7 carbazole degradative plasmids," *Bioscience, Biotechnology and Biochemistry*, vol. 73, no. 3, pp. 744–746, 2009.

[23] P. Deighan, C. Beloin, and C. J. Dorman, "Three-way interactions among the Sfh, StpA and H-NS nucleoid-structuring proteins of *Shigella flexneri* 2a strain 2457T," *Molecular Microbiology*, vol. 48, no. 5, pp. 1401–1416, 2003.

[24] S. C. Dillon, A. D. S. Cameron, K. Hokamp, S. Lucchini, J. C. D. Hinton, and C. J. Dorman, "Genome-wide analysis of the H-NS and Sfh regulatory networks in *Salmonella Typhimurium* identifies a plasmid-encoded transcription silencing mechanism," *Molecular Microbiology*, vol. 76, no. 5, pp. 1250–1265, 2010.

[25] R. T. Dame, M. C. Noom, and G. J. L. Wuite, "Bacterial chromatin organization by H-NS protein unravelled using dual DNA manipulation," *Nature*, vol. 444, no. 7117, pp. 387–390, 2006.

[26] C. Madrid, C. Balsalobre, J. García, and A. Juárez, "The novel Hha/YmoA family of nucleoid-associated proteins: use of structural mimicry to modulate the activity of the H-NS family of proteins," *Molecular Microbiology*, vol. 63, no. 1, pp. 7–14, 2007.

[27] S. de los Rios and J. J. Perona, "Structure of the *Escherichia coli* leucine-responsive regulatory protein Lrp reveals a novel octameric assembly," *Journal of Molecular Biology*, vol. 366, no. 5, pp. 1589–1602, 2007.

SAGE-Hindawi Access to Research
International Journal of Evolutionary Biology
Volume 2011, Article ID 250154, 9 pages
doi:10.4061/2011/250154

Research Article

New Insights on the Evolutionary History of Aphids and Their Primary Endosymbiont *Buchnera aphidicola*

Vicente Pérez-Brocal,[1,2] Rosario Gil,[2,3] Andrés Moya,[1,2,3] and Amparo Latorre[1,2,3]

[1] *Área de Genómica y Salud, Centro Superior de Investigación en Salud Pública (CSISP), Avenida de Cataluña 21, 46020 Valencia, Spain*
[2] *CIBER Epidemiología y Salud Pública (CIBERESP), Spain*
[3] *Departament de Genètica, Institut Cavanilles de Biodiversitat i Biologia Evolutiva, Universitat de València, Apartado Postal 22085, 46071 Valencia, Spain*

Correspondence should be addressed to Vicente Pérez-Brocal, perez_vicbro@gva.es

Received 13 October 2010; Accepted 24 December 2010

Academic Editor: Hiromi Nishida

Since the establishment of the symbiosis between the ancestor of modern aphids and their primary endosymbiont, *Buchnera aphidicola*, insects and bacteria have coevolved. Due to this parallel evolution, the analysis of bacterial genomic features constitutes a useful tool to understand their evolutionary history. Here we report, based on data from *B. aphidicola*, the molecular evolutionary analysis, the phylogenetic relationships among lineages and a comparison of sequence evolutionary rates of symbionts of four aphid species from three subfamilies. Our results support previous hypotheses of divergence of *B. aphidicola* and their host lineages during the early Cretaceous and indicate a closer relationship between subfamilies Eriosomatinae and Lachninae than with the Aphidinae. They also reveal a general evolutionary pattern among strains at the functional level. We also point out the effect of lifecycle and generation time as a possible explanation for the accelerated rate in *B. aphidicola* from the Lachninae.

1. Introduction

Aphids constitute a diversified group of insects widespread and of economical relevance as crop pests. The underlying reason of their ecological success is their novel capability to exploit ecological niches with little competitors, mainly due to their diet based on phloem, which is abundant and of easy access but represents an unbalanced source of nutrients, rich in sugars and poor in amino acids [1]. The clue to the use of new resources lies in the establishment of an obligate endosymbiotic relationship between the ancestor of aphids and a gamma-proteobacterium, the ancestor of *Buchnera aphidicola*. This single event of infection has been dated at least 150–200 million years ago (MYA) [2] according to the fossil record or to 80–150 MYA based on molecular data [3]. As a result of millions of years of cospeciation of host and endosymbiont, the current species of aphids carrying their specific strains of *B. aphidicola* emerged.

The vertical mode of transmission of *B. aphidicola*, from mother to eggs and embryos, together with the location in specific host cells (the bacteriocytes), determines a population scenario for this bacterium characterized by their low effective population size, with frequent bottlenecks and little chance of genetic recombination with other bacteria. As a result, the genome reductive process undergone by *B. aphidicola* encompasses a decrease in the genomic size due to the loss of unnecessary genes in the new intracellular context, the increase in A+T content compared to its free-living relatives, a significant acceleration in evolutionary rates, mainly due to the accumulation of nonsynonymous substitutions, the loss in codon bias, loss of many regulatory proteins and functions, as well as the retention of genes linked to their symbiotic role [4–9].

This particular history of genome reduction is pertinent to understand the coevolution between particular aphid hosts and *B. aphidicola*. Many of the genes that are involved in recombination and/or genetic transference were lost at the beginning of the symbiotic association and, consequently, the *B. aphidicola* clones have evolved independently in each particular host with no or little chance of gene exchange among *B. aphidicola* from different aphid hosts [10].

The comparison of the topology of phylogenetic trees based on aphid genes and those from *B. aphidicola* reveals a perfect match [2, 11]. As a result of this parallel evolutionary pattern, *B. aphidicola* can be regarded as an excellent marker in order to elucidate the evolutionary relationship of aphids harboring particular *B. aphidicola* strains.

The analysis of *B. aphidicola* genes that follow an evolutionary pattern that agrees with the molecular clock hypothesis [12, 13] can be used to estimate the divergence time between pairs of aphids. This is possible because two aphid species, *Acyrthosiphon pisum* and *Schizaphis graminum* belonging to two tribes of the subfamily Aphidinae, have an estimated divergence time calibrated from their fossil record of 50 to 70 MY [14]. In addition, using molecular data from complete *B. aphidicola* genomes available, Pérez-Brocal and coworkers calculated the divergence time of aphids belonging to subfamilies Eriosomatinae (*Baizongia pistaciae*) and Lachninae (*Cinara cedri*) [15]. Based on morphological traits, the subfamilies Eriosomatinae and Lachninae have traditionally been considered very divergent. In fact, most phylogenetic hypotheses based both on morphological and molecular data consider the Lachninae as a sister group of the Aphidinae [11, 16]. However, the position of this subfamily remains controversial, as recent phylogenies based on molecular sequences located the subfamily in a basal position [17–19]. Here, we follow a genomic approach to deepen the evolutionary analyses and propose a phylogeny of the three subfamilies of aphids based on the genome sequence of their primary endosymbionts *B. aphidicola*. In addition, in order to detect if there is any selective effect related to the specific role of the genes, we also gave a closer look to the acceleration pattern of each functional category.

2. Materials and Methods

2.1. Genome Sequences Used in This Study. The genome sequences used in this study were retrieved from GenBank database. The four *B. aphidicola* strains are *B. aphidicola Acyrthosiphon pisum* str. APS (BAp, Accession no. BA000003 [20]), *B. aphidicola Schizaphis graminum* (BSg, Accession no. AE013218 [21]), *B. aphidicola Baizongia pistaciae* (BBp, Accession no. AE016826 [22]), and *B. aphidicola Cinara cedri* (BCc, Accession no. CP000263 [15]). *Escherichia coli* was used as out-group in all comparisons: *E. coli* str. K12 substr. MG1655 (Eco, Accession no. U00096).

2.2. Sequence Alignments. For protein-coding genes, nucleotide sequences were translated into amino acids using the ClustalW tool implemented in the MEGA4 package [23]. The generated amino acid sequences were used, in turn, as a template to align the corresponding nucleotides with MUSCLE v3.6 [24], to reduce ambiguities.

2.3. Estimate of Strain-Specific Evolutionary Rates. *B. aphidicola* BCc was used as a reference strain since it is the one with the lowest gene complement of those analyzed. For each one of the genes present in *B. aphidicola* BCc having an orthologous in at least one of the other *B. aphidicola* strains, an analysis of relative substitution rates between pairs of *B. aphidicola* strains was carried out, using *E. coli* as out-group. Specifically, we applied a Tajima's relative rate test [25] with MEGA4, generating six comparisons for each of the aligned genes. Genes showing accelerated rates were grouped according to a nonredundant categories classification based on that used in the sequencing work on *Aquifex aeolicus* [26], with some modifications [27].

2.4. Estimate of Evolutionary Acceleration among Genomes. The sequence from the 338 protein-coding genes shared by the four *B. aphidicola* strains plus *E. coli* was used to quantify the relative degree of evolutionary acceleration among strains. To do this, nucleotide sequences were concatenated with BioEdit and aligned using the ClustalW tool implemented in the MEGA4 [23]. Three different estimates of substitution rates per site between species i and j (K_{ij}) were carried out with MEGA4, using (a) the total and (b) nonsynonymous nucleotide positions, under the Kimura 2-parameters and the modified Nei-Gojobori methods, respectively, and (c) amino acid sequences, using the JTT substitution matrix. K_{01} and K_{02} were calculated according to Moran [28], being taxon 0 the last common ancestor of the endosymbiont strains compared in each test (taxa 1 and 2). The calculation of total and nonsynonymous substitutions allowed us to account for the phenomenon of saturation. To check for saturation, the "transition and transversion versus divergence" plot was implemented by DAMBE v4.2.13 for the concatenation of shared genes using the first and second positions as well as the third one [29]. This method has been successfully used previously to estimate saturation due to divergence [30–33].

Additionally, for each protein-coding gene under study, the values of both synonymous (d_S) and nonsynonymous (d_N) nucleotide substitutions were calculated, using a modified Nei-Gojobori model (Jukes Cantor) implemented by MEGA4 [23]. To calculate the synonymous (λ_S) and nonsynonymous (λ_N) nucleotide substitutions per million years, we used the expression $\lambda = K/2T$, where K is the number of nucleotide differences per site and T the estimated divergence time. The T values used in these analyses were 107 MY for (*B. aphidicola* BAp-BSg-)BBp, 111 MY for (*B. aphidicola* BAp-BSg-)BCc, and 112 MY for *B. aphidicola* BBp-BCc. These are the previously determined lowest values for each range of estimated divergence times among strains [15], based on the range of 50 to 70 MY since the strains used for calibration (*B. aphidicola* BAp and BSg) diverged as estimated from the fossil record [14]. The global average λ_S and λ_N values for each pair of *B. aphidicola* strains was calculated, as well as the partial average λ_S and λ_N values for each functional category [26, 27] between all the strain pairs.

2.5. Phylogenetic Analyses. Since saturation was achieved at the third position in all comparisons but BAp and BSg, in order to reduce the loss of phylogenetic signal we excluded this position when working with nucleotides to perform our phylogenetic analyses. The concatenated sequence of the 338 protein-coding genes shared by the four *B. aphidicola* strains was used to reconstruct the phylogenetic relationships among them. Maximum Likelihood (ML) analyses were carried out with PAUP4.0b10 [34] for nucleotides, and Phyml_v2.4.5 [35] for amino acids, according to the best models of nucleotide (GTR+I+G) and amino acid (CpREV+I+G+F) substitutions for those genes derived from jModelTest [36] and ProtTest 1.4 [37], respectively. Nucleotides and amino acids were also used for Bayesian analysis, with MrBayes v3.1.2 [38], using four MCMC strands, 1,000,000 generations, with trees sampled every 100 generations. Consensus trees were produced after excluding an initial burn-in of 25% of the samples, as recommended.

In a previous study, the evolutionary analyses of the four *B. aphidicola* strains showed that only 21 genes fulfill the molecular clock hypothesis [15]. The topologies of the 21 phylogenetic trees based on these genes were obtained by ML using PAUP 4.0b10 [34], in order to determine the most plausible evolutionary relationships among strains and compare them with the phylogenetic reconstruction.

2.6. Statistical Analyses. All statistical analyses were performed using the software package R (http://www.r-project.org) [39]. A chi-square analysis was applied to the global distribution of the accelerated genes among *B. aphidicola* strains compared to the distribution within functional families, to test whether any particular functional category contains a significantly increased or reduced number of accelerated genes. Twelve comparisons with Yates' correction were carried out, at a significance level $\alpha = 0.05$.

The average rates of synonymous (λ_S) and nonsynonymous (λ_N) substitutions per site per million years of the six possible comparisons among *B. aphidicola* strains were compared using a one-way ANOVA analysis followed by Tukey's range tests to find which means are significantly different from one another.

3. Results

3.1. Comparison of the Evolutionary Rates in B. aphidicola Strains at a Genome Level. The relative rate test on the 338 concatenated protein-coding genes (Table 1) reveals that, since the last common ancestor of each pair of strains, the accumulation of both nucleotide and amino acid substitutions, as well as the nonsynonymous substitution rates follows different rates in the different strains, but the values obtained using all three parameters are equivalents for any given strain pair. Thus, for the nucleotide sequences, a similar pattern of relative evolutionary rates was observed when total and nonsynonymous substitution rates are considered. *B. aphidicola* BSg and BAp show a similar rate (1.12 : 1), the one in *B. aphidicola* BBp being slightly higher (1.3-1.4-fold that of *B. aphidicola* BSg and BAp) and *B. aphidicola* BCc being the one with more accelerated rates (1.7-fold that of *B. aphidicola* BBp and more than 2-fold that of *B. aphidicola* BAp and BSg). As for the amino acid sequences, the relative acceleration shows a similar patter as the one observed for the nucleotides, but with values in *B. aphidicola* BCc of 2 to 3-fold those of *B. aphidicola* BBp and BAp-BSg, respectively.

The evolutionary acceleration among genomes was also determined through the analysis of the synonymous (λ_S) and nonsynonymous (λ_N) nucleotide substitutions per million years. The results show that both rates exhibit an opposite pattern (Figure 1). Differences in both λ_S and λ_N are statistically significant (ANOVA test, significance level 0.05), clustering into three separate groups for λ_S and two groups for λ_N, according to Tukey's range tests. When synonymous substitutions (Figure 1(a)) are considered, the more accelerated rate is found in the comparison between strains *B. aphidicola* BAp and BSg, a second group includes *B. aphidicola* BBp with the two aforementioned, and the least accelerated one includes all rates in which *B. aphidicola* BCc is involved. A different pattern is found for nonsynonymous substitutions (Figure 1(b)), where the more accelerated group includes all the comparisons involving *B. aphidicola* BCc, and the other one includes the remaining three comparisons.

3.2. Analyses of the Evolutionary Rates at a Functional Level. The general pattern identified at a genomic level is reproduced at every functional category (see Section 2), with the same three and two groups of *B. aphidicola* strain pairs found in λ_S and λ_N, respectively (Figure 2). On the other hand, no significant differences are found among functional categories in any strain for λ_S (Figure 2(a)). However, a significant increase in λ_N is found for the genes involved in cell envelope in all the strains ($P < .05$) and to a lesser extent in the category of poorly characterized genes (Figure 2(b)). This could be due to a significant acceleration of the flagellar genes still remaining in *B. aphidicola*, especially in BCc, the strain which has undergone the most drastic reduction in the flagellar machinery.

In a previous study, we determined the global relative distribution of accelerated genes displayed by the strains, using Tajima's relative rate test [15]. According to this test, *B. aphidicola* BCc presents a higher number of accelerated genes (56%–83%), while *B. aphidicola* BBp presents intermediate values (0.6%–35%), and the fewest appear in *B. aphidicola* BSg and specially BAp. This trait is observed in each functional category with no significant differences (Figure 3). This homogeneous distribution of the accelerated genes across functional categories was tested by the application of χ^2 tests, based on the observed number of accelerated genes for each category and the expected number of genes based on the totality of them for each pair of strains. None of the tests was statistically significant at $P < .05$ (Table 2).

3.3. Phylogenetic Analyses Show an Evolutionary Radiation Pattern. According to the molecular clock hypothesis, two taxa sharing a common ancestor should have accumulated the same number of substitutions since they diverge. In the *B. aphidicola* case, only 21 genes do not reject the molecular clock hypothesis [15]. These genes can be used to identify the phylogenetic relationships among the strains under study, which will also reflect the relationships among their insect hosts. However, three different tree topologies appear in

Table 1: Relative rate tests for the 338 concatenated protein-coding genes shared by the four *B. aphidicola* strains included in this study plus *E. coli* [a]: (a) nonsynonymous sites, (b) all nucleotides, and (c) amino acids.

(a)

Taxon 1	Taxon 2	Taxon 3	$\overline{K}_{12}$	$\overline{K}_{13}$	$\overline{K}_{23}$	$\overline{K}_{13} - \overline{K}_{23}$	$\overline{K}_{01}/\overline{K}_{02}$
BAp	BSg	Eco	0.152	0.339	0.348	−0.009	0.89
BAp	BBp	Eco	0.319	0.339	0.395	−0.056	0.70
BAp	BCc	Eco	0.380	0.339	0.494	−0.155	0.42
BSg	BBp	Eco	0.319	0.348	0.395	−0.047	0.74
BSg	BCc	Eco	0.377	0.348	0.494	−0.146	0.44
BBp	BCc	Eco	0.392	0.395	0.494	−0.099	0.60

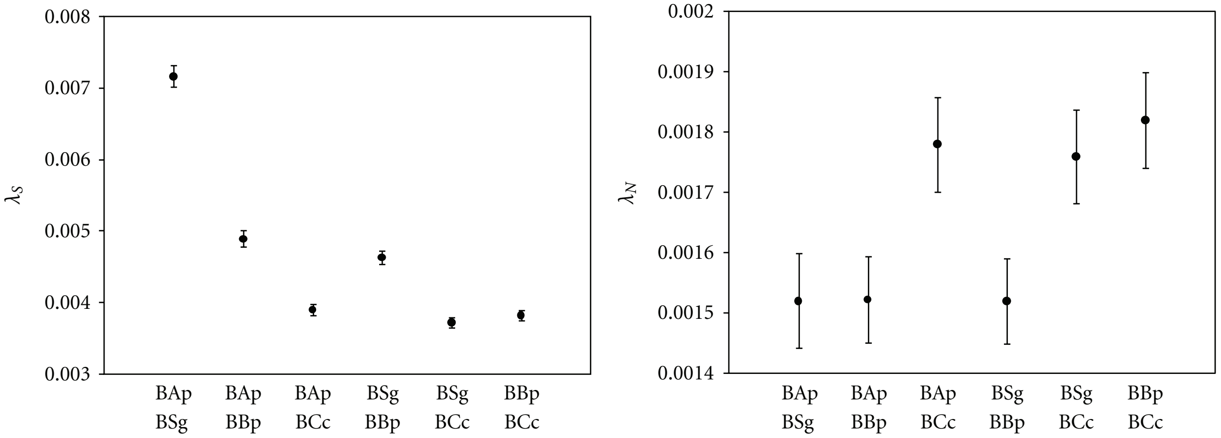
(b)

Taxon 1	Taxon 2	Taxon 3	$\overline{K}_{12}$	$\overline{K}_{13}$	$\overline{K}_{23}$	$\overline{K}_{13} - \overline{K}_{23}$	$\overline{K}_{01}/\overline{K}_{02}$
BAp	BSg	Eco	0.242	0.617	0.630	−0.013	0.89
BAp	BBp	Eco	0.421	0.617	0.685	−0.068	0.72
BAp	BCc	Eco	0.452	0.617	0.791	−0.174	0.50
BSg	BBp	Eco	0.417	0.63	0.685	−0.055	0.77
BSg	BCc	Eco	0.445	0.63	0.791	−0.161	0.47
BBp	BCc	Eco	0.463	0.685	0.791	−0.106	0.62

(c)

Taxon 1	Taxon 2	Taxon 3	$\overline{K}_{12}$	$\overline{K}_{13}$	$\overline{K}_{23}$	$\overline{K}_{13} - \overline{K}_{23}$	$\overline{K}_{01}/\overline{K}_{02}$
BAp	BSg	Eco	0.350	0.814	0.842	−0.028	0.85
BAp	BBp	Eco	0.845	0.814	1.001	−0.187	0.64
BAp	BCc	Eco	1.126	0.814	1.410	−0.596	0.30
BSg	BBp	Eco	0.850	0.842	1.001	−0.159	0.68
BSg	BCc	Eco	1.180	0.842	1.410	−0.568	0.33
BBp	BCc	Eco	1.186	1.001	1.410	−0.409	0.48

[a] In each test, taxa 1 and 2 represent *B. aphidicola* strains, taxon 3 represents *E. coli*, and taxon 0 represents the last common ancestor of taxa 1 and 2. $\overline{K}_{ij}$ is the estimate of substitutions per site between taxon i and taxon j.



(a) B. aphidicola strains

(b) B. aphidicola strains

Figure 1: Global average values (and confidence interval of 95%) of (a) synonymous (λ_S) and (b) nonsynonymous (λ_N) nucleotide substitutions per site per million years. The divergence times among strains are 50 (BAp-BSg), 107 (BAp-BBp and BSg-BBp), 111 (BAp-BCc and BSg-BCc), and 112 (BBp-BCc) MY, respectively. The numbers of shared protein-coding genes are 348 (BAp-BSg), 347 (BAp-BBp), 354 (BAp-BCc), 343 (BSg-BBp), 350 (BSg-BCc), and 350 (BBp-BCc), respectively.

Table 2: Yates' chi-square tests for the accelerated genes classified by functional category in four *B. aphidicola* strains compared in pairs. Acceleration is based on Tajima's relative rate tests. The total number of comparisons for each particular category and pair of strains is shown in brackets ( ). A/B: number of accelerated genes in A compared to B and in B compared to A, respectively.

Observed	Pairs of strains					
Functional category	BAp/BSg	BAp/BBp	BAp/BCc	BSg/BBp	BSg/BCc	BBp/BCc
(1) Information storage and processing	5/12 (160)	0/54 (160)	0/137 (162)	0/47 (158)	0/122 (160)	1/86 (160)
(2) Protein processing, folding, and secretion	1/1 (25)	0/11 (24)	0/19 (25)	0/10 (24)	0/18 (25)	0/14 (24)
(3) Cellular processes	0/0 (10)	0/5 (10)	0/7 (10)	0/3 (10)	0/7 (10)	1/7 (10)
(4) Metabolism	2/9 (103)	3/34 (103)	0/86 (104)	2/32 (104)	0/88 (105)	0/61 (106)
(5) Cell envelope	0/0 (14)	1/7 (13)	0/12 (14)	1/7 (13)	0/12 (14)	0/10 (13)
(6) Poorly characterized	1/1 (33)	1/10 (34)	0/29 (35)	0/7 (32)	0/23 (33)	0/15 (34)
Total	9/23 (345)	5/121 (344)	0/290 (350)	3/106 (341)	0/270 (347)	2/193 (347)
Expected	Pairs of strains					
Functional category	BAp/BSg	BAp/BBp	BAp/BCc	BSg/BBp	BSg/BCc	BBp/BCc
(1) Information storage and processing	4.17/10.67	2.33/56.28	0.00/134.23	1.39/49.11	0.00/124.50	0.92/88.99
(2) Protein processing, folding and secretion	0.65/1.67	0.35/8.44	0.00/20.71	0.21/7.46	0.00/19.45	0.14/13.35
(3) Cellular processes	0.26/0.67	0.15/3.52	0.00/8.29	0.09/3.11	0.00/7.78	0.06/5.56
(4) Metabolism	2.69/6.87	1.50/36.23	0.00/86.17	0.91/32.33	0.00/81.70	0.61/58.96
(5) Cell envelope	0.36/0.93	0.19/4.57	0.00/11.60	0.11/4.04	0.00/10.89	0.07/7.23
(6) Poorly characterized	0.86/2.20	0.49/11.96	0.00/29.00	0.28/9.95	0.00/25.68	0.20/18.91
$\chi^2 s$ (with Yates' correction, 5 d.f.)	0.501/0.933	3.491/1.908	0.000/0.195	4.776/2.762	0.000/0.720	7.455/1.598
P value	.992/.968	.625/.862	1.000/.999	.444/.737	1.000/.982	.189/.902

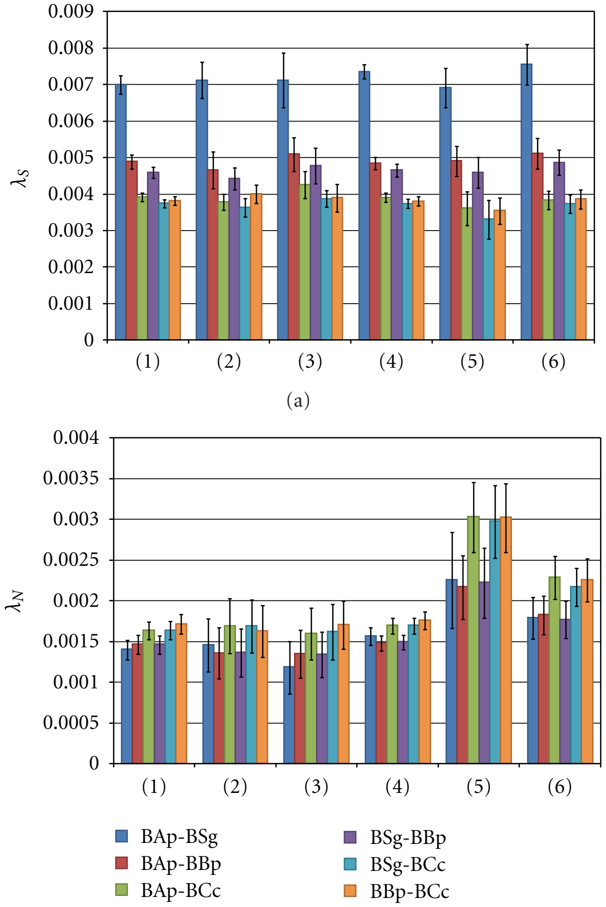
a similar number of cases for these 21 genes (see Figure 4). Six genes generated the topology a (*B. aphidicola* BCc basal), seven the topology b (*B. aphidicola* BBp basal), and eight the topology c (*B. aphidicola* BCc and BBp clustered). Therefore, the analysis of these genes, individually considered, does not resolve the position of the *B. aphidicola* BCc and BBp strains. This result points at the possibility of a radiation within a relatively short period of time, giving rise to the subfamilies. To confirm this point, and in order to solve the deepest relationship among subfamilies, a more exhaustive phylogenetic reconstruction was carried out, based on all the concatenated protein-coding genes shared by the four *B. aphidicola* strains. The resulting phylogenetic tree (Figure 5) shows the same topology as tree c in Figure 4, that is, a well supported clade consisting of both members of the subfamily Aphidinae, as expected, and another clade that shows a clustering of *B. aphidicola* BBp and BCc, also with the maximum statistical support. The uneven branch length, being that of *B. aphidicola* BCc significantly longer, indicates the evolutionary acceleration experienced by this strain. The topology obtained using amino acid sequences is identical, but the relative length of *B. aphidicola* BCc's branch is even longer, reflecting a higher value of nonsynonymous substitutions.

4. Discussion

4.1. Reconstruction of the Evolutionary History of Aphids Belonging to Subfamilies Aphidinae, Eriosomatinae and Lachninae. Aphids emerged as a monophyletic group of viviparous insects about 250 MYA as a divergent group from the oviparous Adelgidae and Phylloxeridae [11]. The basal radiation of the family Aphididae was dated by molecular data to the Cretaceous, 80 to 150 MYA [3]. Although the initial development of aphids took place on gymnosperms during the Mesozoic, most of their current diversity is linked to angiosperms, especially to grass [40]. The extraordinary diversity of aphids found today, affecting specially the subfamily Aphididae, started during the Tertiary (Miocene), as a consequence of the proliferation of herbaceous angiosperms [41, 42].

The phylogenetic position of the subfamily Lachninae within the Aphididae is controversial. Traditionally, phylogenies based on both morphological characters [11, 16] and on mitochondrial rDNA [3] have placed them as a monophyletic group clustering with the Aphidinae. However, phylogenies based on sequences from both nuclear and mitochondrial aphid genes (long-wavelength opsin gene, the elongation factor 1α gene, and mitochondrial genes encoding ATPase 6 subunit and the subunit II of the cytochrome oxidase), as well as those based on their primary endosymbiont *B. aphidicola* (16S rDNA and the β subunit of the F-ATPase complex) [17–19] place them as a basal group apart from the Aphidinae. This fact has implications about those aphids feeding on conifers (such as most members of the subfamily Lachninae, including *C. cedri*) being regarded as ancestral to groups feeding on angiosperms or, alternatively, as more recent secondarily derived conifer suckers.

Our phylogenetic analysis supports the presence of one clade clustering *B. aphidicola* BBp and BCc, and another



(a)

(b)

Figure 2: Average values (and confidence interval of 95%) of (a) synonymous (λ_S) and (b) nonsynonymous (λ_N) nucleotide substitutions per site per million years for each functional category. The numbers of shared protein-coding genes are 348 (BAp-BSg), 347 (BAp-BBp), 354 (BAp-BCc), 343 (BSg-BBp), 350 (BSg-BCc), and 350 (BBp-BCc), respectively. (1) Information storage and processing; (2) protein processing, folding, and secretion; (3) cellular processes; (4) metabolism; (5) cell envelope; (6) poorly characterized. Each given comparison is colored as illustrated above.

clade consisting of *B. aphidicola* BAp and BSg. This result is consistent with a panorama of a rapid evolutionary radiation of the main subfamilies of aphids, during the early Cretaceous (144-100 MYA), which seems concordant with previous proposals [3]. In addition, our evolutionary molecular data from *B. aphidicola* point out that aphids belonging to subfamilies Eriosomatinae and Lachninae share a common ancestor more closely related than compared to the members of subfamily Aphidinae. If true, our data refute the traditional phylogenetic reconstructions that placed Aphidinae and Lachninae as a monophyletic group [11]. However, we do not have evidence to conclude whether, within the subfamily Lachninae, tribes feeding on conifers are ancestral or more recent than those living on herbaceous angiosperms, since our analysis does not resolve which strain (and thus which host aphid) is basal compared to the others. To solve this point, it would be necessary to sequence the genome of a greater number of *B. aphidicola* strains, including members of the different tribes from the subfamily Lachninae (work in progress). This would allow us to establish the date of divergence between those tribes and, thus, try to relate this fact to the change of vegetal host in either direction.

4.2. Accelerated Evolutionary Rates in B. aphidicola within the Subfamily Lachninae. From an evolutionary perspective, the protein-coding genes of *B. aphidicola* show higher ratios of nonsynonymous versus synonymous substitutions (d_N/ d_S) than those of free-living bacteria, due to an accelerated rate of nonsynonymous substitutions, a characteristic of bacterial endosymbionts [14, 28], where mutations with amino acid replacement are not efficiently eliminated by a relaxed purifying selection, leading to a greater accumulation of amino acid changes than in free-living bacteria. These nonsynonymous substitutions end up in fixation by genetic drift, due to the mode of transmission and the population dynamics of *B. aphidicola*. This acceleration of evolutionary rates is particularly evident in *B. aphidicola* BCc, presumably because factors promoting the accumulation of nonsynonymous substitutions are more intense in this strain. One of those factors is the extreme reduction of the repair machinery, barely able to counterbalance the accumulation of slightly deleterious mutations. In addition, there is a stronger effect of genetic drift that promotes the fixation of slightly deleterious mutations probably imposed by its coexistence within the aphid with a secondary symbiont, *Serratia symbiotica*, and its larger size compared to other *B. aphidicola* lineages [43]. A closer look at the particular genes that contribute to this acceleration observed in *B. aphidicola* BCc allows us to conclude that they are distributed among different functional categories, with none of them accumulating significant differences in the proportion of accelerated genes (as seen in Figure 3 and Table 2). This fact reveals that the process of gene degradation acts on any type of gene independently of their functional role. However, our results indicate that even if the accelerated genes are scattered homogeneously across all the functional categories in all *B. aphidicola* strains, genes of some functional categories, such as cellular envelope, are significantly more accelerated within all the lineages. That points to the ongoing action of selective constraints affecting nonsynonymous substitution rates.

Regarding synonymous substitutions, when pairs of strains of *B. aphidicola* were compared based on the average number of synonymous substitutions per site (d_S), a greater accumulation was observed in the *B. aphidicola* BBp strain compared to bacteria from aphids of the subfamily Aphidinae (*B. aphidicola* BAp and BSg), while the smallest value is found between the *B. aphidicola* BAp and BSg strains [15]. However, if the temporary factor is considered, the rates of synonymous nucleotide substitutions per site and million years are greater in the endosymbionts from the Aphidinae (*B. aphidicola* BAp and BSg strains), registering the *B. aphidicola* BCc strain the smallest values. These results demonstrate that the synonymous substitution rate in *B. aphidicola* is a variable character, yet the explanation for

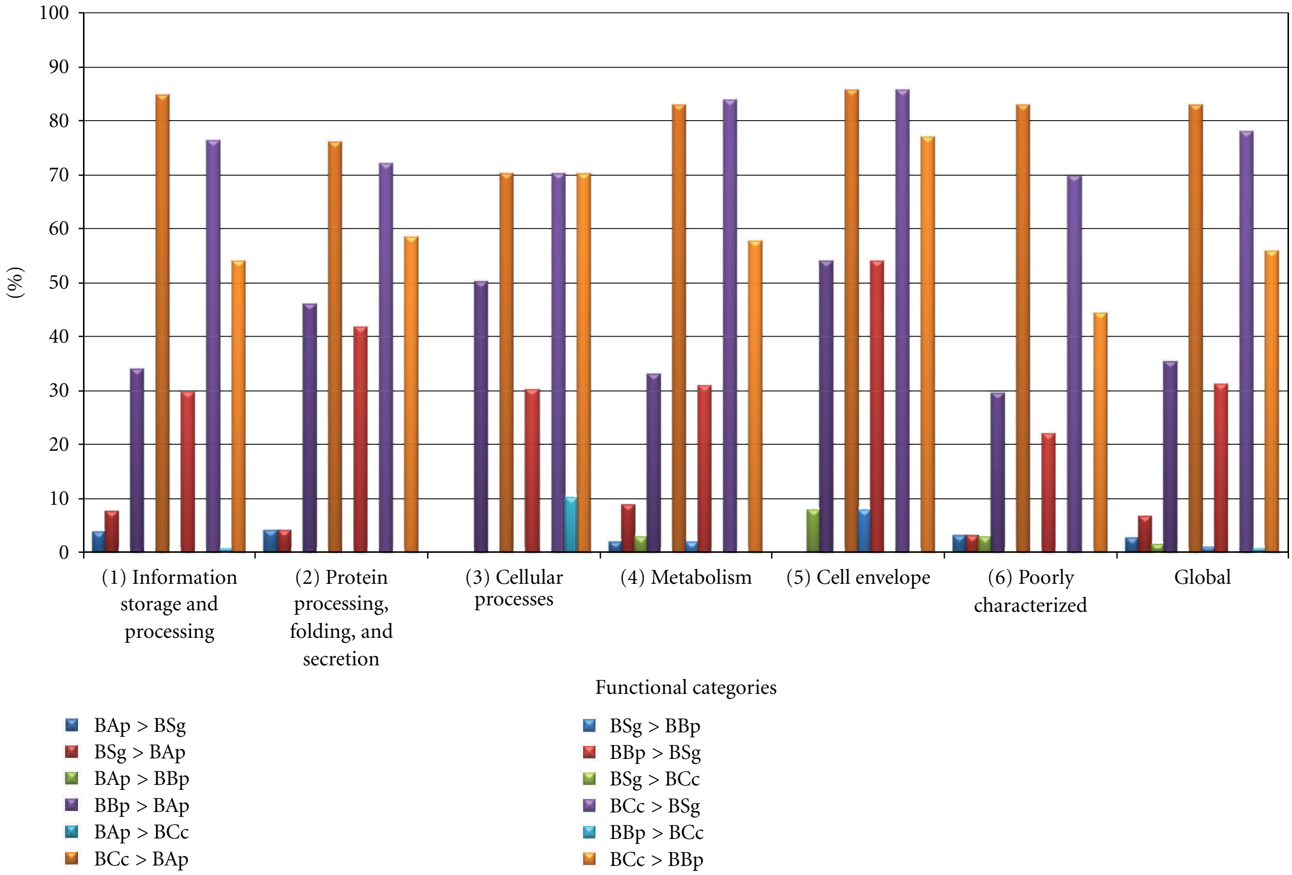
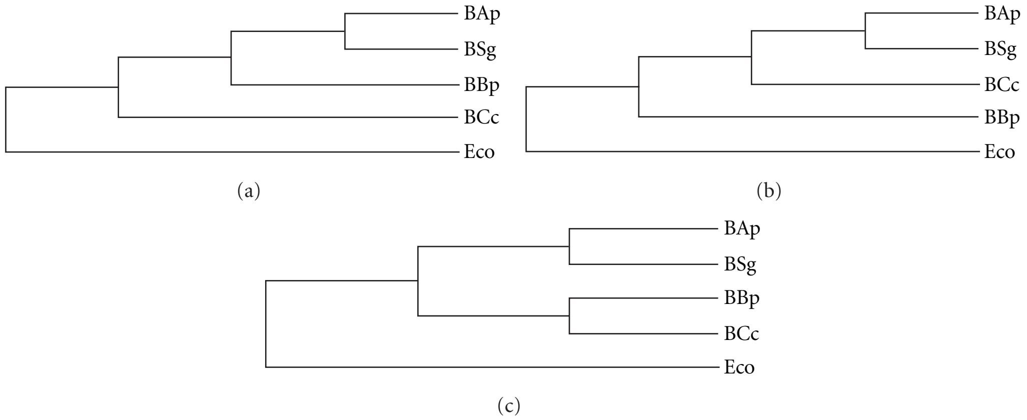


Figure 3: Relative distribution of the accelerated genes based on their functional category, between pairs of *B. aphidicola* strains. Accelerated genes were calculated by Tajima's relative rate tests. A > B indicates a significantly higher accumulation of substitutions in strain A than in strain B ($P < .05$).



(a)

(b)

(c)

Figure 4: Topologies of the phylogenetic trees for the 21 genes that follow the hypothesis of molecular clock [15]. The trees were obtained by maximum likelihood, with the program PAUP 4.0b10.

these divergent patterns is not obvious. As stated elsewhere [7, 14, 44], these differences can be attributed to differences in the host's life cycle, as well as ecological factors such as host-alternation and variations in the effective population size showed by the two members of the Aphidinae subfamily compared to the other two aphid lineages. Additionally a differential mutation rate per generation cannot be ruled out. For example, endosymbionts from aphids with short generation times can accumulate more synonymous mutations per million years (case of the Aphidinae) than those

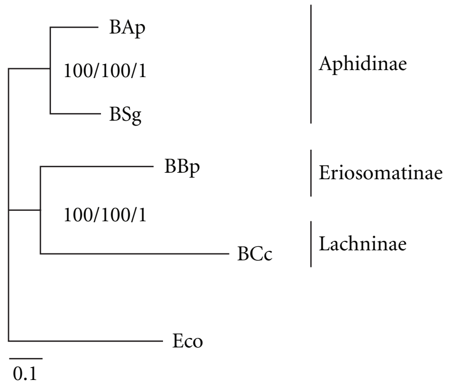


Figure 5: Phylogenetic tree obtained by maximum likelihood using PAUP4.0b10 on nucleotide sequences and the GTR+I+G evolutionary model. Topologies obtained from amino acid sequences, using Phyml_v2.4.5 and MrBayes v3.1.2, are identical. Trees are based on the concatenated sequence of the 338 protein-coding genes shared by the four *B. aphidicola* strains and *E. coli*. Numbers beside the internal nodes are the maximum likelihood bootstrap values from 300 resamplings obtained with PAUP4.0b10, Phyml and the Bayesian MCMC posterior probability, respectively. The scale bar represents the number of nucleotide substitutions per site.

with longer generation times, such as the Eriosomatinae and the Lachninae. Future studies are required to understand the evolutionary processes driving these patterns.

Acknowledgments

Financial support was provided by Grant BFU2009-12895-C02-01/BMC (Ministerio de Educación y Ciencia, Spain) to A. Latorre and European Community's Seventh Framework Programme (FP7/2007–2013) under Grant Agreement number 212894 and Prometeo/2009/092 (Conselleria d'Educació, Generalitat Valenciana, Spain) to A. Moya.

References

[1] J. Sandstrom and J. Pettersson, "Amino acid composition of phloem sap and the relation to intraspecific variation in pea aphid (*Acyrthosiphon pisum*) performance," *Journal of Insect Physiology*, vol. 40, no. 11, pp. 947–955, 1994.

[2] N. A. Moran, M. A. Munson, P. Baumann, and H. Ishikawa, "A molecular clock in endosymbiotic bacteria is calibrated using the insect hosts," *Proceedings of the Royal Society B*, vol. 253, no. 1337, pp. 167–171, 1993.

[3] C. D. von Dohlen and N. A. Moran, "Molecular data support a rapid radiation of aphids in the Cretaceous and multiple origins of host alternation," *Biological Journal of the Linnean Society*, vol. 71, no. 4, pp. 689–717, 2000.

[4] J. J. Wernegreen and N. A. Moran, "Evidence for genetic drift in endosymbionts (*Buchnera*): analyses of protein-coding genes," *Molecular Biology and Evolution*, vol. 16, no. 1, pp. 83–97, 1999.

[5] L. Klasson and S. G. E. Andersson, "Evolution of minimal-gene-sets in host-dependent bacteria," *Trends in Microbiology*, vol. 12, no. 1, pp. 37–43, 2004.

[6] J. J. Wernegreen, A. O. Richardson, and N. A. Moran, "Parallel acceleration of evolutionary rates in symbiont genes underlying host nutrition," *Molecular Phylogenetics and Evolution*, vol. 19, no. 3, pp. 479–485, 2001.

[7] T. Itoh, W. Martin, and M. Nei, "Acceleration of genomic evolution caused by enhanced mutation rate in endocellular symbionts," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, no. 20, pp. 12944–12948, 2002.

[8] J. J. Wernegreen, "Genome evolution in bacterial endosymbionts of insects," *Nature Reviews Genetics*, vol. 3, no. 11, pp. 850–861, 2002.

[9] A. Mira and N. A. Moran, "Estimating population size and transmission bottlenecks in maternally transmitted endosymbiotic bacteria," *Microbial Ecology*, vol. 44, no. 2, pp. 137–143, 2002.

[10] The International Aphid Genomics Consortium, "Genome aequence of the pea aphid *Acyrthosiphon pisum*," *PLoS Biology*, vol. 8, no. 2, Article ID e1000313, 2010.

[11] O. E. Heie, "Palaeontology and phylogeny," in *Aphids: Their Biology, Natural Enemies and Control*, A. K. Minks and P. Harrewijn, Eds., vol. 2A, pp. 367–391, Elsevier, Amsterdam, The Netherlands, 1987.

[12] E. Zuckerkandl and L. Pauling, "Molecular disease, evolution, and genetic heterogeneity," in *Horizons in Biochemistry*, M. Kasha and B. Pullman, Eds., pp. 189–225, Academic Press, New York, NY, USA, 1962.

[13] E. Zuckerkandl and L. Pauling, "Evolutionary divergence and convergence in proteins," in *Evolving Genes and Proteins*, V. Bryson and H. J. Vogel, Eds., pp. 97–166, Academic Press, New York, NY, USA, 1965.

[14] M. A. Clark, N. A. Moran, and P. Baumann, "Sequence evolution in bacterial endosymbionts having extreme base compositions," *Molecular Biology and Evolution*, vol. 16, no. 11, pp. 1586–1598, 1999.

[15] V. Pérez-Brocal, R. Gil, S. Ramos et al., "A small microbial genome: the end of a long symbiotic relationship?" *Science*, vol. 314, no. 5797, pp. 312–313, 2006.

[16] W. Wojciechowski, *Studies on the Systematic System of Aphids (Homoptera, Aphidinea)*, Uniwersytet Slaski, Katowice, Poland, 1992.

[17] D. Martinez-Torres, C. Buades, A. Latorre, and A. Moya, "Molecular systematics of aphids and their primary endosymbionts," *Molecular Phylogenetics and Evolution*, vol. 20, no. 3, pp. 437–449, 2001.

[18] B. Ortiz-Rivas, A. Moya, and D. Martínez-Torres, "Molecular systematics of aphids (Homoptera: Aphididae): new insights from the long-wavelength opsin gene," *Molecular Phylogenetics and Evolution*, vol. 30, no. 1, pp. 24–37, 2004.

[19] B. Ortiz-Rivas and D. Martínez-Torres, "Combination of molecular data support the existence of three main lineages in the phylogeny of aphids (Hemiptera: Aphididae) and the basal position of the subfamily Lachninae," *Molecular Phylogenetics and Evolution*, vol. 55, no. 1, pp. 305–317, 2010.

[20] S. Shigenobu, H. Watanabe, M. Hattori, Y. Sakaki, and H. Ishikawa, "Genome sequence of the endocellular bacterial symbiont of aphids *Buchnera* sp. APS," *Nature*, vol. 407, no. 6800, pp. 81–86, 2000.

[21] I. Tamas, L. Klasson, B. Canbäck et al., "50 million years of genomic stasis in endosymbiotic bacteria," *Science*, vol. 296, no. 5577, pp. 2376–2379, 2002.

[22] R. C. H. J. van Ham, J. Kamerbeek, C. Palacios et al., "Reductive genome evolution in *Buchnera aphidicola*," *Proceedings*

of the National Academy of Sciences of the United States of America, vol. 100, no. 2, pp. 581–586, 2003.

[23] K. Tamura, J. Dudley, M. Nei, and S. Kumar, "MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0," *Molecular Biology and Evolution*, vol. 24, no. 8, pp. 1596–1599, 2007.

[24] R. C. Edgar, "MUSCLE: multiple sequence alignment with high accuracy and high throughput," *Nucleic Acids Research*, vol. 32, no. 5, pp. 1792–1797, 2004.

[25] F. Tajima, "Simple methods for testing the molecular evolutionary clock hypothesis," *Genetics*, vol. 135, no. 2, pp. 599–607, 1993.

[26] G. Deckert, P. V. Warren, T. Gaasterland et al., "The complete genome of the hyperthermophilic bacterium *Aquifex aeolicus*," *Nature*, vol. 392, no. 6674, pp. 353–358, 1998.

[27] R. Gil, F. J. Silva, E. Zientz et al., "The genome sequence of Blochmannia floridanus: comparative analysis of reduced genomes," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, no. 16, pp. 9388–9393, 2003.

[28] N. A. Moran, "Accelerated evolution and Muller's rachet in endosymbiotic bacteria," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 93, no. 7, pp. 2873–2878, 1996.

[29] X. Xia and Z. Xie, "DAMBE: software package for data analysis in molecular biology and evolution," *Journal of Heredity*, vol. 92, no. 4, pp. 371–373, 2001.

[30] A. T. Marques, A. Antunes, P. A. Fernandes, and M. J. Ramos, "Comparative evolutionary genomics of the HADH2 gene encoding Aβ-binding alcohol dehydrogenase/17β-hydroxysteroid dehydrogenase type 10 (ABAD/HSD10)," *BMC Genomics*, vol. 7, article 202, 2006.

[31] M. G. Fain and P. Houde, "Multilocus perspectives on the monophyly and phylogeny of the order Charadriiformes (Aves)," *BMC Evolutionary Biology*, vol. 7, article 35, 2007.

[32] M. Farfán, D. Miñana-Galbis, M. C. Fusté, and J. G. Lorén, "Divergent evolution and purifying selection of the flaA gene sequences in *Aeromonas*," *Biology Direct*, vol. 4, article 23, 2009.

[33] M. Daly, L. C. Gusmão, A. J. Reft, and E. Rodríguez, "Phylogenetic signal in mitochondrial and nuclear markers in sea anemones (cnidaria, Actiniaria)," *Integrative and Comparative Biology*, vol. 50, no. 3, pp. 371–388, 2010.

[34] D. L. Swofford, *PAUP∗. Phylogenetic analysis using parsimony (∗and other methods). Version 4*, Sinauer Associates, Sunderland, Mass, USA, 2002.

[35] S. Guindon and O. Gascuel, "A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood," *Systematic Biology*, vol. 52, no. 5, pp. 696–704, 2003.

[36] D. Posada, "jModelTest: phylogenetic model averaging," *Molecular Biology and Evolution*, vol. 25, no. 7, pp. 1253–1256, 2008.

[37] F. Abascal, R. Zardoya, and D. Posada, "ProtTest: selection of best-fit models of protein evolution," *Bioinformatics*, vol. 21, no. 9, pp. 2104–2105, 2005.

[38] J. P. Huelsenbeck and F. Ronquist, "MRBAYES: Bayesian inference of phylogenetic trees," *Bioinformatics*, vol. 17, no. 8, pp. 754–755, 2001.

[39] R Development Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2010, http://www.R-project.org.

[40] V. F. Eastop, "Biotypes of aphids," in *Perspectives in Applied Biology*, A. D. Lowe, Ed., vol. 51 of *Bulletin of the Entomological Society of New Zealand*, pp. 40–51, 1973.

[41] O. E. Heie, "Aphid ecology in the past and a new view on the evolution of Macrosiphini," in *Individuals, Populations and Patterns in Ecology*, S. R. Leather, A. D. Watt, N. J. Mills, and K. F. A. Walters, Eds., pp. 409–418, Intercept, Andover, UK, 1994.

[42] O. E. Heie, "The evolutionary history of aphids and a hypothesis on the coevolution of aphids and plants," *Bollettino di Zoologia Agraria e di Bachicoltura*, vol. 28, pp. 149–155, 1996.

[43] L. Gómez-Valero, A. Latorre, and F. J. Silva, "The evolutionary fate of nonfunctional DNA in the bacterial endosymbiont Buchnera aphidicola," *Molecular Biology and Evolution*, vol. 21, no. 11, pp. 2172–2181, 2004.

[44] H. Ochman, S. Elwyn, and N. A. Moran, "Calibrating bacterial evolution," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 96, no. 22, pp. 12638–12643, 1999.

SAGE-Hindawi Access to Research
International Journal of Evolutionary Biology
Volume 2011, Article ID 781642, 10 pages
doi:10.4061/2011/781642

Research Article

Parallel Evolution and Horizontal Gene Transfer of the *pst* Operon in *Firmicutes* from Oligotrophic Environments

Alejandra Moreno-Letelier,[1,2] Gabriela Olmedo,[2] Luis E. Eguiarte,[1] Leon Martinez-Castilla,[3] and Valeria Souza[1]

[1] *Departamento de Ecologia Evolutiva, Instituto de Ecologia, Universidad Nacional Autónoma de México, Apdo. Postal 70-275, Ciudad Universitaria, 04510 México D. F., Mexico*
[2] *Departamento de Ingeniería Genética, CINVESTAV Campus Guanajuato, Apdo. Postal 629, 36500 Irapuato, Mexico*
[3] *Departamento de Bioquimica, Facultad de Quimica, Universidad Nacional Autonoma de México, Apdo. Postal 70-275, Ciudad Universitaria, 04510 México D. F., Mexico*

Correspondence should be addressed to Valeria Souza, souza@servidor.unam.mx

Received 22 October 2010; Accepted 22 December 2010

Academic Editor: Hiromi Nishida

The high affinity phosphate transport system *(pst)* is crucial for phosphate uptake in oligotrophic environments. Cuatro Cienegas Basin (CCB) has extremely low P levels and its endemic *Bacillus* are closely related to oligotrophic marine *Firmicutes*. Thus, we expected the *pst* operon of CCB to share the same evolutionary history and protein similarity to marine *Firmicutes*. Orthologs of the *pst* operon were searched in 55 genomes of *Firmicutes* and 13 outgroups. Phylogenetic reconstructions were performed for the *pst* operon and 14 concatenated housekeeping genes using maximum likelihood methods. Conserved domains and 3D structures of the phosphate-binding protein (PstS) were also analyzed. The *pst* operon of *Firmicutes* shows two highly divergent clades with no correlation to the type of habitat nor a phylogenetic congruence, suggesting horizontal gene transfer. Despite sequence divergence, the PstS protein had a similar 3D structure, which could be due to parallel evolution after horizontal gene transfer events.

1. Introduction

Phosphorus is an essential nutrient for multiple processes such as the synthesis of DNA, RNA, ATP, and many other pathways involving phosphorylation [1]. However, it is not an abundant element on the planet and can only be obtained form organic detritus or from tectonics and volcanism [2, 3], so, its availability is a limiting factor for all life forms. As growth rate and primary productivity are highly dependent on phosphorus [4–6], bacteria have different mechanisms for the uptake and storage of phosphates to be able to cope with this limitation [1, 7–9].

Some of the genes involved in phosphorus metabolism belong to the *pho* regulon that is induced by phosphorus starvation by a two-component regulatory system in several bacteria such as *Escherichia coli*, *Bacillus subtilis,* and *Cyanobacteria* [8, 10–13]. The *pho* regulon is comprised of 20 or so genes that include phosphatases, phosphate transport systems, and other enzymes used to assimilate phosphorus form other sources such as phosphonates [8]. Even though the *pho* regulon is found in both Eubacteria and Archaea, the number and identity of the genes are highly variable and not always congruent with the 16S rRNA gene phylogeny of the organisms [11, 14]. It is also to be expected that the genes involved in phosphate uptake and metabolism would be under strong selection.

Among the genes of the *pho* regulon, the high affinity phosphate transport system (*pst*) is thought to be responsible for phosphate uptake under nutrient stress [8, 10]. *Pst* is a typical ABC transport system encoded in 4 to 6 genes in a single operon [10, 15–17]. As an ABC transporter, the *pst* operon belongs to one of the largest gene families and is found in all Eubacteria and Archaea and the level of sequence divergence indicates an ancient origin of each lineage of transporters [18, 19].

The genes of the *pst* operon are arranged in the following way: the *pstS* gene, coding for a periplasmic protein that binds phosphate with high affinity; *pstC* and *pstA,* coding for the two proteins proposed to form the inner membrane channel; *pstB,* coding for an ATPase that energizes the transport [18]. However, some variation exists in the number of genes in the operon. In *Escherichia coli* and *Clostridium acetobutylicum*, the gene *phoU*, coding for a repressor of the *pho* regulon, is also located in the operon [15, 17], while in *B. subtilis* and its close relatives there are no *phoU* orthologs. Also, the gene *pstB* is duplicated (*pstBA* and *pstBB*; [10]). The *pst* operon presents further variation in *Cyanobacteria*, where the genes *pstS* or *pstB* may be missing from the operon depending on the strain and environmental conditions [11], or additional *pstS* copies may be present although not associated to the operon [11, 20].

The *pst* phosphate uptake system is particularly crucial in oligotrophic environments such as the North Pacific, North Atlantic and the Eastern Mediterranean Sea [6, 21]. Metagenomic studies have shown that there are some functional adaptations for P uptake in such oligotrophic waters [7, 8, 20, 22]. Another example of an extreme oligotrophic environment is the Cuatro Cienegas Basin (CCB), that presents very low levels of P in the ecosystem [4, 23, 24]. Phosphate concentrations range from 0.008 to 0.6 μM, in Pozas Azules and Rio Mezquites, respectively (E. Rebollar and F. García-Oliva pers. com.; [4]), but for most water systems P concentrations lie below the threshold concentration for the expression of the *pho* regulon in *B. subtilis* (0.1 mM; [10]).

CCB is an isolated oasis in the center of the Chihuahuan Desert, with water systems rich in microbial mats and stromatolites, and its microbiota exhibits ancestral marine affinities [9, 25–29]. Despite the extreme oligotrophy of the ecosystem, CCB has a high level of diversity and species endemism both at the macro- and microscopic levels [24, 25, 30–33]. We believe that this high rate of diversification is a consequence of the extreme oligotrophy of the ecosystem [24], where the lack of available P promotes both reproductive and geographic isolation, by limiting replication and the frequency of genetic exchange [24, 34–36]. Moreover, two of the newly sequenced taxa, *Bacillus coahuilensis* and *Bacillus sp. m3–13*, have particular adaptations to low P environments. Unlike *Escherichia coli* or *B. subtilis*, CCB and marine *Bacillus* lack the low affinity phosphate uptake system so they must rely on the high affinity transport system [9, 27].

There are some comparative studies about genes involved in phosphorus uptake in *Cyanobacteria* [8, 11, 20], but as far as we know, no studies exist in other bacterial groups. Hence, we believe that an analysis of the phosphorus uptake in the *Firmicutes* from CCB in comparison to sequenced *Firmicutes* from different environments could help us understand the evolution of the high affinity phosphate transport system. *Firmicutes* is a cosmopolitan and ancient lineage [37], and their diversification happened during a time in the Earth's history where P was very scarce [3, 5]. We expected the *pst* operon of the *Firmicutes* from CCB to have a marine affinity and to be related (both in sequence and structure) to the *pst* operons of other marine *Firmicutes* that live in oligotrophic waters.

In this study we analyzed for the first time the evolutionary relationships, gene architecture, of the *pst* operons of 55 complete genomes of the main lineages of *Firmicutes* [38] with special emphasis on CCB and marine taxa, as well as the protein structure of PstS from a few *Bacillus*. To evaluate phylogenetic congruity between the phosphate uptake genes and housekeeping genes, expected to reflect vertical descent, we performed a phylogenetic reconstruction of the genes of the *pst* operon and of 14 proteins of the core genome of *Firmicutes*. We also compared the protein structure of phosphate-binding protein PstS of *Bacillus* from oligotrophic and eutrophic environments, to try to evaluate any association between protein sequence and structure to the environment in which the members of *Firmicutes* live.

2. Materials and Methods

2.1. Phylogenetic Reconstructions. We used the amino acid sequence of the substrate-binding protein gene *pstS* of *Bacillus coahuilensis* and *Bacillus subtilis* [9, 39] to identify the orthologs of the *pst* operon in the draft and complete genomes of the main lineages of *Firmicutes* (for accession numbers see Table S1 of Supplementary Material available online at doi: 10.4061/2011/781642). Searches were performed using psi-Blast, and the sequences identified with at least 30% of identity over a minimum of 70% in length, and *e*-value $<10^{-35}$ were considered as orthologs [27]. As the *pstS* gene can be duplicated in some genomes, the operon structure of the genes in the operon was manually checked in all cases, and only genes with the highest bit scores and lowest *e*-values were considered in cases of multiple hits in the same genome. For the cases with multiple hits of the entire operon, all those extra copies of the operon were also included in the analysis. We also included 11 sequences of the *pst* operon of non-*Firmicutes* that had Blast scores better than our threshold. As outgroups, we included 2 genomes of non-*Firmicutes*: *Thermotoga maritima* (*Thermotogae*) and *Pelobacter carbinolicus* (*δ-Proteobacteria*), that gave the next best hitting scores below our threshold. Due to the high sequence variation of the *pstS* gene and the different number of copies of the other genes in the operon, the reconstruction was done with genes *pstC*, *pstA*, and *pstB* in a concatenated matrix. The *phoU* gene was excluded because it was missing in several *Bacillus* species. When both *pstBA* and *pstBB* were present, *pstBB* was used in the analysis as it is the ortholog of *pstB*; *pstBA* was not included, The *pstBB* gene was identified on the basis of genomic context, as it is the second gene coding for an atp-binding protein in the operon. We compared the topology of the *pst* operon phylogeny with a phylogeny reconstructed from 14 concatenated amino acid sequences from genes from the core genome of *Firmicutes*. Those 14 genes were chosen from a set of genes already identified by Maughan [38] and Alcaraz et al. ([27]; GI from *B. subtilis*: 2632976, 2632269, 2632399, 2634021, 2636597, 16079910, 50812244, 50812227, 16079600, 16077523, 16080084, 16077081, and 16077661, 2635239). Four *Cyanobacteria*, *Chloroflexus aurantiacus* and *Thermotoga maritima* were used as outgroups (list of strain names and genome accession numbers in Table S1 of supplementary material). To establish a temporal frame of events, we dated the 14 gene phylogeny using a penalized likelihood method implemented by r8s [40]. The calibration of the tree was done using dates of geological events: the divergence of aerobic firmicutes was fixed at 2300 million years ago, a conservative date for the Great Oxidation Event [3, 37]. The divergence of CCB *Firmicutes* and their closest relatives was constrained to have a minimum age of 35 my, that corresponds to the uplift of the Sierra Madre Oriental that finally isolated CCB from the Gulf of Mexico [41].

All reconstructions were done using amino acid sequences aligned using MUSCLE [42] and a Maximum Likelihood approach, implemented by Raxml v.7.0.4 [43]; (CIPRES portal: http://www.phylo.org/) with a LG substitution model chosen using ProtTest v 2.1 [44] with the Akaike Information Criterion, 4 substitution categories, and allowing Raxml to estimate the proportion of invariant sites and the gamma shape parameter. For both datasets 100 bootstrap replicates were performed.

2.2. PstS Protein Motifs and 3D Structure. The main conserved motifs of the substrate-binding protein PstS were detected using the MEME suite ([45]; http://meme.sdsc.edu/meme/) using the default parameters and an alignment of the PstS amino acid sequence from all the *Firmicutes*, but including in this alignment only one sequence representative of the *Bacillus cereus* group.

The 3D structure of the PstS protein was modeled based on the 3-D structures of PstS from *Yersinia pestis* (PDB ID: 2z22) and PstS-1 from *Mycobacterium tuberculosis* (PDB ID: 1pc3; [46]). Only the PstS from *B. subtilis*, *B. coahuilensis*, *B. sp. m3-13,* and *B. sp. NRRL-14911* were modeled using the web-based module of MODELLER using the default settings (http://modbase.compbio.ucsf.edu/ModWeb20-html/modweb.html; [47]). Comparisons of 3D models were performed with TOPOFIT. This method only takes into account the geometric attributes of the proteins and not the sequence similarity, so it is able to find structure homology in highly variable proteins [48]. The quality of the models was evaluated with the r.m.s.d. value (root mean squared deviation) and the z-score (a measure of the energy separation between two protein folds). Images were prepared with CHIMERA (http://www.cgl.ucsf.edu/chimera).

3. Results

Using a psi-Blast search we were able to find orthologs of the *pst* operon in all members of *Firmicutes*, several *Cyanobacteria* and *Archaea* as well as in some *Bacteroidetes, Fusobacteria, Actinobacteria,* and *Planctomycetes*. In the search we observed that the gene architecture of the operon showed variation within and outside *Firmicutes* (Figure 1). Several taxa had a duplication in tandem of the gene *pstB,* as was the case for *Bacillus subtilis*, *Listeria monocytogenes,* and *Streptococcus pneumoniae*. *Cyanobacteria* generally lacked the gene *phoU* in the same operon, and it was missing in the *B. subtilis* group, and *Clostridium tetani*. *Synechococcus sp. 7002* also lacked the *pstB* gene in the operon (Figure 1). Duplications of the *pstS* gene were found in the *Bacillus cereus* group, *Exiguobacterium* spp., *Brevibacillus brevis, Bacillus sp. B14905*, *Enterococcus faecalis*, *Lactobacillus plantarum,* and *Geobacillus kaustophilus*, but the entire operon was only duplicated in *Streptococcus pneumoniae* and *Symbiobacterium thermophilus.* In the *B. cereus* group, an incomplete copy (*pstSCA*, lacking *pstB*) of the *pst* operon was found, similar to that of *B. subilis*, thus it was not used for the concatenated phylogeny, but only for the PstS phylogeny (see Figure S1 in the supplementary material). All *Bacillus* from CCB and most marine *Bacillus* had just one copy of the *pstS* gene (the exception, *Bacillus sp. B14905*).

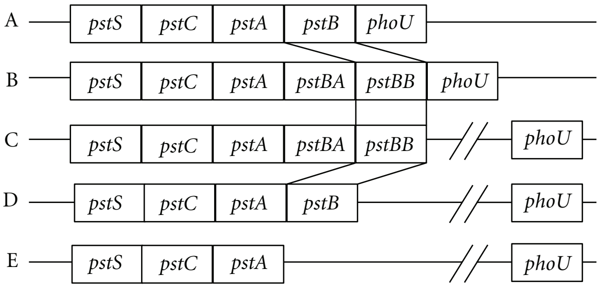


Figure 1: Gene architecture of the *pst* operon in different groups of bacteria. A: *Bacillus cereus* group, marine *Bacillus* and *Clostridium;* B: *Listeria* and *Streptococcus;* C: *Bacillus subtilis group, Bacillus marisflavi,* and *Bacillus sp. CH108;* D: *Brevibacillus, Oceanobacillus, Desulfitobacterium, Acaryochloris*; E: *Synechococcus, Geobacillus kaustophilus, Sebaldella termitidis, B. cereus* group. In C–E, the regulatory gene *phoU* is found elsewhere in the genome, not as part of the operon or entirely missing.

The phylogenetic reconstruction of the concatenated PstC, PstA, and PstB protein sequences showed two distinct and highly supported clades (Figure 2) that bear no relation to either the type of habitat or the phylogenetic relationships obtained with the amino acid sequences from housekeeping genes (Figure 3). Reconstructions made with each sequence independently, yielded the same basic topology, with minor differences in branch length (data not shown), so the phylogenetic signal was present in all three genes. We named one of the clades "*cereus*-like", which consists of the *pstSCABU* operon structure (operon architecture A, in Figure 1), and it includes all members of the *B. cereus* group, most of *Bacillus* and *Staphylococcus*, *Exiguobacterium,* an anaerobic soil firmicute *Desulfitobacterium hafniense*, and most noteworthy, several *Cyanobacteria* and *Archaea* (Figure 2). None of the members of that clade have the duplication of gene *pstB*, and only two taxa (*Desulfitobacterium* and *Oceanobacillus*) lack the gene *phoU* in the operon (for the operon structure of all taxa in the dataset see Table S1 of Supplementary Material).

The other highly supported clade was named "*subtilis*-like" (operon architecture C, in Figure 1) and it included the members of the *B. subtilis* group, a marine *Bacillus*, *Bacillus marisflavi* and its sister species *Bacillus sp. CH108* from CCB, *Listeria*, *Clostridium*, some host-associated firmicutes, and

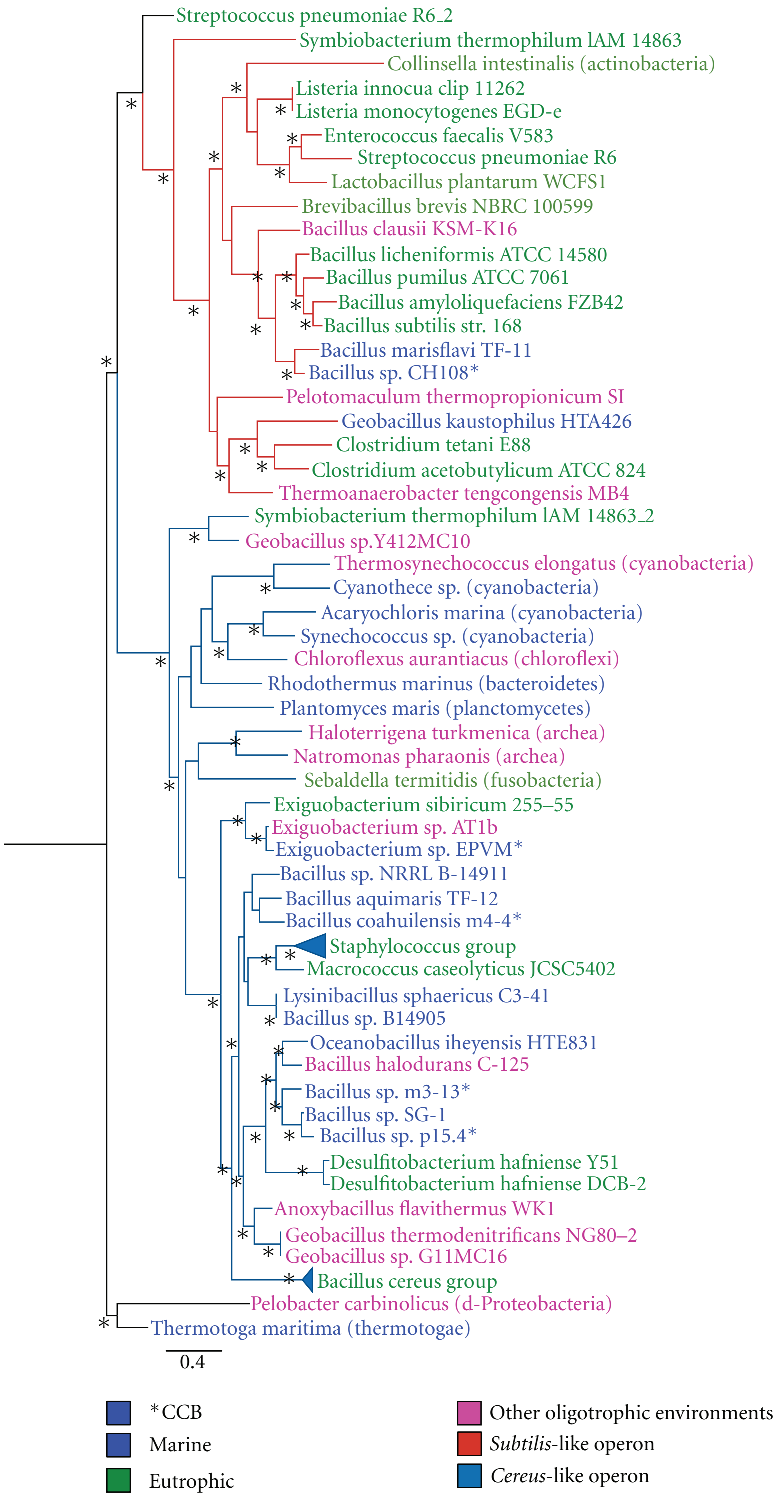

Streptococcus pneumoniae R6_2
Symbiobacterium thermophilum lAM 14863
Collinsella intestinalis (actinobacteria)
Listeria innocua clip 11262
Listeria monocytogenes EGD-e
Enterococcus faecalis V583
Streptococcus pneumoniae R6
Lactobacillus plantarum WCFS1
Brevibacillus brevis NBRC 100599
Bacillus clausii KSM-K16
Bacillus licheniformis ATCC 14580
Bacillus pumilus ATCC 7061
Bacillus amyloliquefaciens FZB42
Bacillus subtilis str. 168
Bacillus marisflavi TF-11
Bacillus sp. CH108*
Pelotomaculum thermopropionicum SI
Geobacillus kaustophilus HTA426
Clostridium tetani E88
Clostridium acetobutylicum ATCC 824
Thermoanaerobacter tengcongensis MB4
Symbiobacterium thermophilum lAM 14863_2
Geobacillus sp.Y412MC10
Thermosynechococcus elongatus (cyanobacteria)
Cyanothece sp. (cyanobacteria)
Acaryochloris marina (cyanobacteria)
Synechococcus sp. (cyanobacteria)
Chloroflexus aurantiacus (chloroflexi)
Rhodothermus marinus (bacteroidetes)
Plantomyces maris (planctomycetes)
Haloterrigena turkmenica (archea)
Natromonas pharaonis (archea)
Sebaldella termitidis (fusobacteria)
Exiguobacterium sibiricum 255–55
Exiguobacterium sp. AT1b
Exiguobacterium sp. EPVM*
Bacillus sp. NRRL B-14911
Bacillus aquimaris TF-12
Bacillus coahuilensis m4-4*
Staphylococcus group
Macrococcus caseolyticus JCSC5402
Lysinibacillus sphaericus C3-41
Bacillus sp. B14905
Oceanobacillus iheyensis HTE831
Bacillus halodurans C-125
Bacillus sp. m3-13*
Bacillus sp. SG-1
Bacillus sp. p15.4*
Desulfitobacterium hafniense Y51
Desulfitobacterium hafniense DCB-2
Anoxybacillus flavithermus WK1
Geobacillus thermodenitrificans NG80–2
Geobacillus sp. G11MC16
Bacillus cereus group
Pelobacter carbinolicus (d-Proteobacteria)
Thermotoga maritima (thermotogae)
0.4
*CCB
Marine
Eutrophic
Other oligotrophic environments
Subtilis-like operon
Cereus-like operon

Figure 2: Maximum likelihood phylogenetic reconstruction of the concatenated PstC, PstA, *and* PstB (PstBB) protein sequences encoded by the *pst* operon. Branch colors indicate the two divergent clades: *subtilis*-like and *cereus*-like. Tag colors indicate the type of habitat where each species is found. Bootstrap values above 70% are indicated with an asterisk. The phylogeny of the individual proteins has a very similar topology (data not shown).

an *Actinobacteria* (*Collinsella intestinalis;* Figure 2). The gene architecture of the operon in the members of this clade is more variable. The members of the genus *Bacillus* have the gene *pstB* duplicated and lack the *phoU* gene in the operon or entirely (Figure 1). *Listeria*, *Enterococcus* and *Streptococcus* also have the *pstB* gene duplicated but the gene *phoU* is in the operon, and although the *pst* operon in *Clostridium* has an architecture similar to that of *B. cereus,* it is very different at sequence level, as seen from the fact that these two are located in different clades (Figure 2).

The high variation at the amino acid sequence level observed for PstS is common for substrate-binding proteins of ABC transporters [18]. In our case, PstS had a shape parameter of the Gamma distribution for site rates of 3.4745, while the PstC, PstA, and PstB proteins had a shape parameter of 1.0559 and the proteins used for the housekeeping gene phylogeny had a shape parameter of 0.7002 (Mega 4; [49]).

The *pst* operon was not monophyletic for marine *Bacillus*, even though marine *Bacillus* are mostly monophyletic, as determined from house-keeping genes (Figure 3) and from other reconstructions (Figure 3; [27]). The main incongruence observed in the tree obtained from the amino acid sequence of the proteins encoded in the *pst* operon is the position of the *B. subtilis* group in a clade with the *Listeria* and *Streptococcus* sequences, instead of grouping with the rest of *Bacillus* taxa (Figure 2). This contrasts with the house-keeping genes phylogeny, were *B. subtilis* and its close relatives are found well within the *Bacillus* clade (Figure 3). Also, the *B. marisflavi-B.sp* CH108 clade, that groups with other marine *Bacillus* in the house-keeping genes phylogeny (Figure 3), appears as a sister group of *B. subtilis* and close taxa in the *pst* operon reconstruction (Figure 2). Also, *Bacillus sp. m3-13* from CCB, appears within the *B. subtilis* clade in the house-keeping genes phylogeny, but is sister to *Bacillus sp. SG-1* from the Gulf of Mexico in the *pst* phylogeny. Another main topological incongruence of the *pst* phylogeny compared to the one done with housekeeping genes, is that of sister taxa *Bacillus halodurans,* and *Bacillus clausii* that are found in different clades: *B. clausii* is found in a clade with *B. subtilis*, *B. marisflavi,* and *Bacillus sp. CH108* while *B. halodurans* forms a monophiyletic clade with some marine *Bacillus* and in turn, is sister to the clade of *Desulfitobacterium hafniense*, an anaerobic species that is found in a basal position in the *Firmicutes* clade obtained from housekeeping genes (Figure 3).

Regarding the encountered motifs of protein PstS (Figure 4), we observed a marked difference between the PstS of the *cereus*-group and that from the *subtilis*-group. Motifs 4 and 5 are located in the same region of the protein but are markedly different, while motif 3 is found in both sets of sequences, but in the *subtilis*-like PstS it had a lower *e*-value (Figure 4). Despite the marked difference at the sequence level, the PstS proteins of *B. subtilis, B. sp.* m3-13, and *B. sp.* NRRL-14911, the 3D structures of the proteins showed similarity with geometry-based alignments (low r.m.s.d. and high z-scores; Figures 5(a)–5(c)), with the exception of *B. coahuilensis* that showed the worse fit values of the three comparisons (Figure 5). This could be a product of the initial 3D model based on a more distantly-related PstS, because the PstS from *B. coahuilensis* also had bad fitting rmsd and z-score values with the PstS of *B. sp.* NRRL-14911 despite having high sequence similarity (74% identity). Thus, the 3D model of the PstS of *B. coahuilensis* can still be improved.

Despite the structure similarities among the PstS of *B. subtilis*, *B. sp.* m3-13, and *B. sp.* NRRL-14911, the active site showed some striking differences in amino acid composition. *B. subtilis* and most of the taxa that grouped in the *subtilis*-like clade have an arginine as the first residue of the active site, just like *Y. pestis* and *M. tuberculosis*, while the firmicutes of the *cereus*-like clade have a proline in the same position (Figure 5(d)). Also, some members of the *cereus*-like clade, like *B. sp.* m3-13, *B. halodurans, Desulfitobacterium spp., O. iheyensis, B. sp.* SG-1, and the *Staphylococcus* group had also a histidine in the second residue of the active site, while all the other taxa had a serine (Figure 5(d)). In view of these changes in amino acid composition an additional codon-based *Z*-test of selection was made for the *pstS* gene with Mega 4 (Tables S2 and S3 of supplementary material; [49]). In all cases, dS (synonymous substitutions) was significantly higher than dN (non-synonymous substitutions), suggesting purifying selection.

4. Discussion

As expected, the *pst* operon was found in all *Firmicutes*. However, not so expected was the finding of two types of operons in these Gram positives. We describe them as *subtilis-like* and *cereus*-like operons, after the best known members of *Bacillus*. Even more interesting, these operons were not shared by descent in the monophyletic groups of *Bacillus*, neither they were operons related to the particular habitat of the strains. Both operons were very divergent from each other at the amino acid sequence level, suggesting independent parallel evolution.

The high divergence of the two types of *pst*-operons in *Firmicutes* and their incongruence with species phylogeny is most noteworthy. Contrary to what was expected, the *pst* operon of marine *Bacillus* is not monophyletic, even though marine and CCB *Bacillus* are resolved as a monophyletic group in the 14 housekeeping gene reconstruction. Therefore, we cannot argue for a common origin due to shared environmental conditions. The patchy distribution of either type of operon within the phylogeny of *Firmicutes* suggests horizontal gene transfer, especially considering closely related species with entirely different *pst* operons (*B. clausii* and *B. halodurans*, *B. subtilis* clade, and *B. sp m3-13*; Figure 2). Another possible explanation of these divergent operons, would be an ancient duplication. However, it is a hypothesis difficult to test, since only very few taxa have both kinds of operons or at least partial copies. At least in the case of the *B. cereus* group, the *subtilis*-like *pstS* copy is more closely related to that of *Clostridium*, while the *pstS* of *B. subtilis* is closely related to *Listeria*, suggesting an independent acquisition (see Figure S1 in supplementary material). If this partial operon or any of the extra copies of *pstS* found in different organisms are functional and

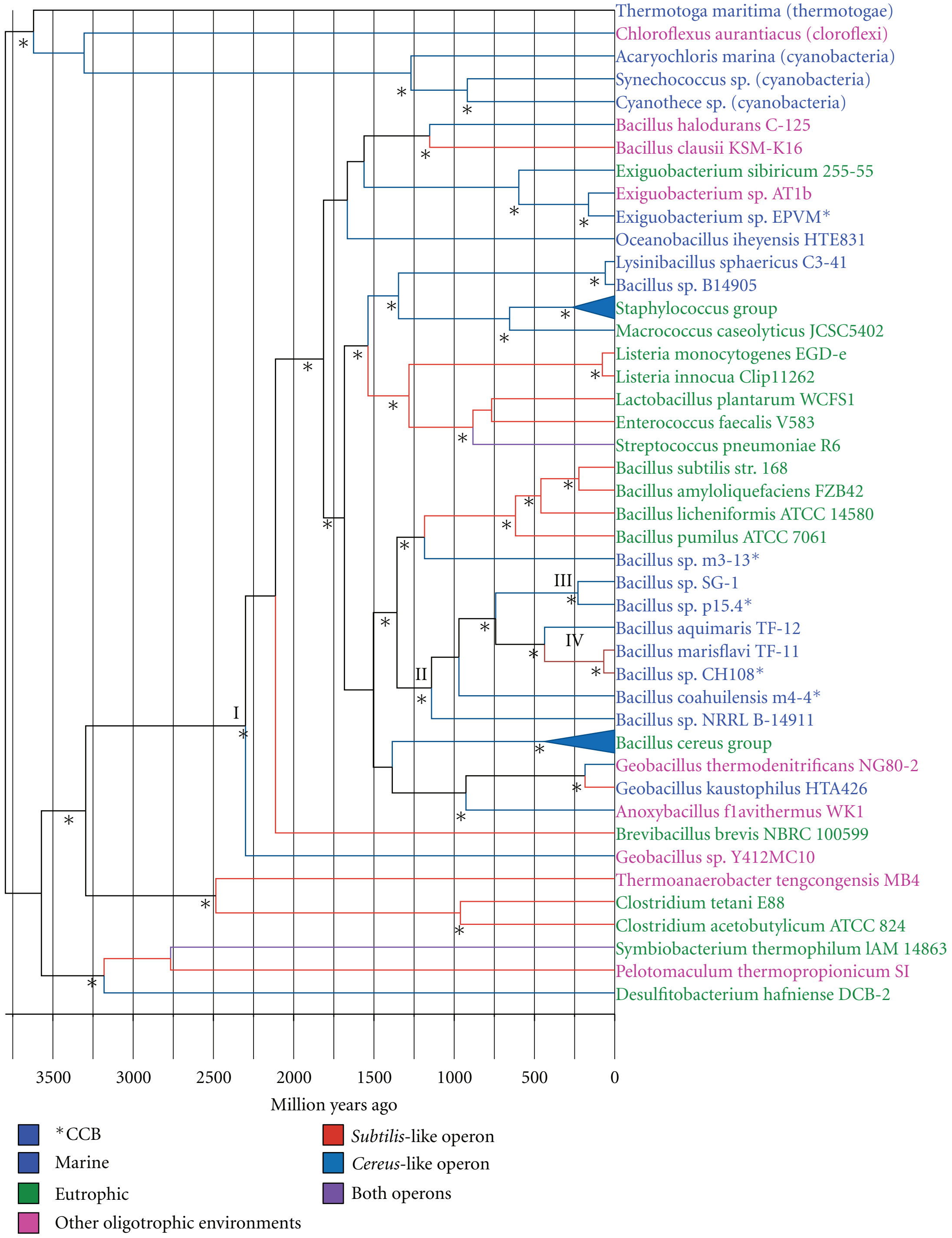

Thermotoga maritima (thermotogae)
Chloroflexus aurantiacus (cloroflexi)
Acaryochloris marina (cyanobacteria)
Synechococcus sp. (cyanobacteria)
Cyanothece sp. (cyanobacteria)
Bacillus halodurans C-125
Bacillus clausii KSM-K16
Exiguobacterium sibiricum 255-55
Exiguobacterium sp. AT1b
Exiguobacterium sp. EPVM*
Oceanobacillus iheyensis HTE831
Lysinibacillus sphaericus C3-41
Bacillus sp. B14905
Staphylococcus group
Macrococcus caseolyticus JCSC5402
Listeria monocytogenes EGD-e
Listeria innocua Clip11262
Lactobacillus plantarum WCFS1
Enterococcus faecalis V583
Streptococcus pneumoniae R6
Bacillus subtilis str. 168
Bacillus amyloliquefaciens FZB42
Bacillus licheniformis ATCC 14580
Bacillus pumilus ATCC 7061
Bacillus sp. m3-13*
III
Bacillus sp. SG-1
Bacillus sp. p15.4*
IV
Bacillus aquimaris TF-12
Bacillus marisflavi TF-11
II
Bacillus sp. CH108*
Bacillus coahuilensis m4-4*
I
Bacillus sp. NRRL B-14911
Bacillus cereus group
Geobacillus thermodenitrificans NG80-2
Geobacillus kaustophilus HTA426
Anoxybacillus f1avithermus WK1
Brevibacillus brevis NBRC 100599
Geobacillus sp. Y412MC10
Thermoanaerobacter tengcongensis MB4
Clostridium tetani E88
Clostridium acetobutylicum ATCC 824
Symbiobacterium thermophilum lAM 14863
Pelotomaculum thermopropionicum SI
Desulfitobacterium hafniense DCB-2
3500
3000
2500
2000
1500
1000
500
0
Million years ago
*CCB
Marine
Eutrophic
Other oligotrophic environments
Subtilis-like operon
Cereus-like operon
Both operons

Figure 3: Maximum likelihood phylogeny of *Firmicutes* based on the concatenated amino acid sequence of 14 housekeeping genes and dated with a penalized likelihood method. The branch colors indicate the type of operon present in each taxon. Tag colors refer to the type of habitat. Clade I corresponds to aerobic *Firmicutes* and clade II includes CCB and marine *Bacillus.* Clade I had a fixed age of 2300 my and clades III and IV had a fixed minimum age of 35 my. Bootstrap values above 70% are denoted with an asterisk. Clades with branch lengths of 0 were collapsed (*D. hafniense DCB-2-D. hafniense Y51* and *G. thermodenitrificans NG80-2-G.sp. G11MC16*).

expressed, is still not known and would require experimental validation.

The relative conservation of gene architecture in *Firmicutes* as opposed to what is seen in Cyanobacteria suggests fewer rearrangements due to phages or some sort of constraint for the transcription and/or regulation of the *pst* operon that has kept the gene architecture fairly constant since the divergence of *Cyanobacteria* and *Firmicutes* around

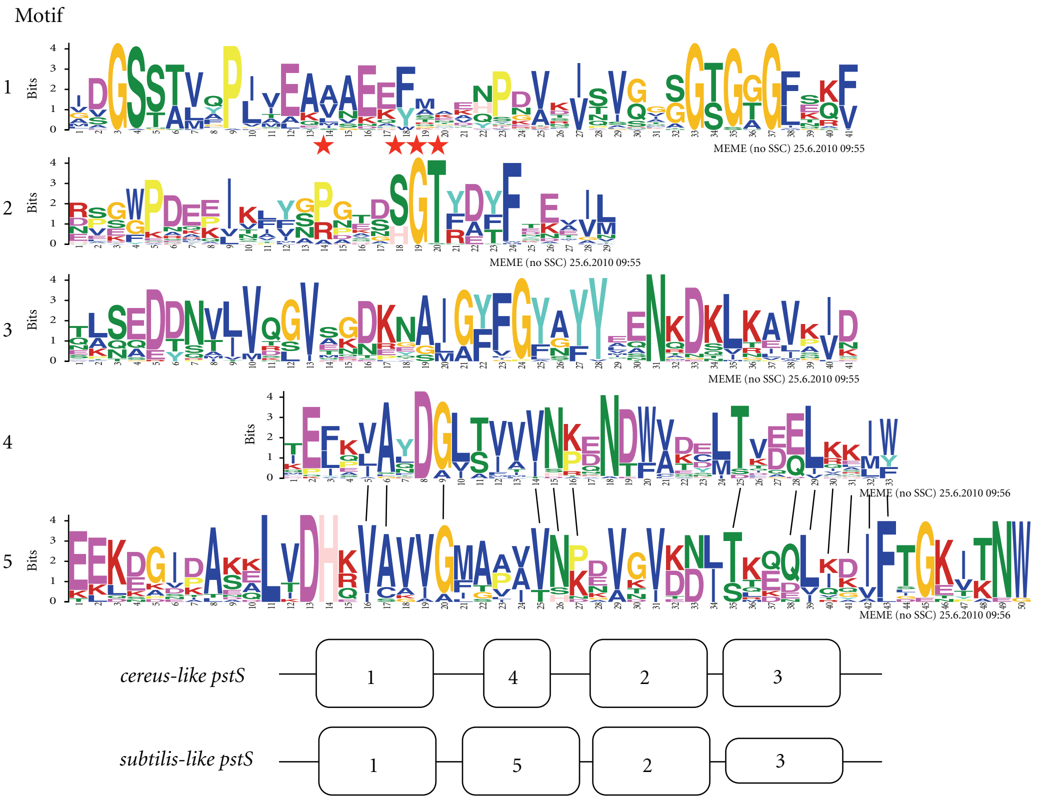


Figure 4: Conserved motifs of the PstS protein for both *cereus*-like and *subtilis*-like operons. Red stars on top of residues on motif 2 indicate the binding site of phosphate. Motifs 4 and 5 are aligned (black lines) to show homologous amino acid positions. The height of the blocks is proportional to the *e*-value of each motif. Motif 3 for *subtilis*-like PstS has an *e*-value < 10^{-10}.

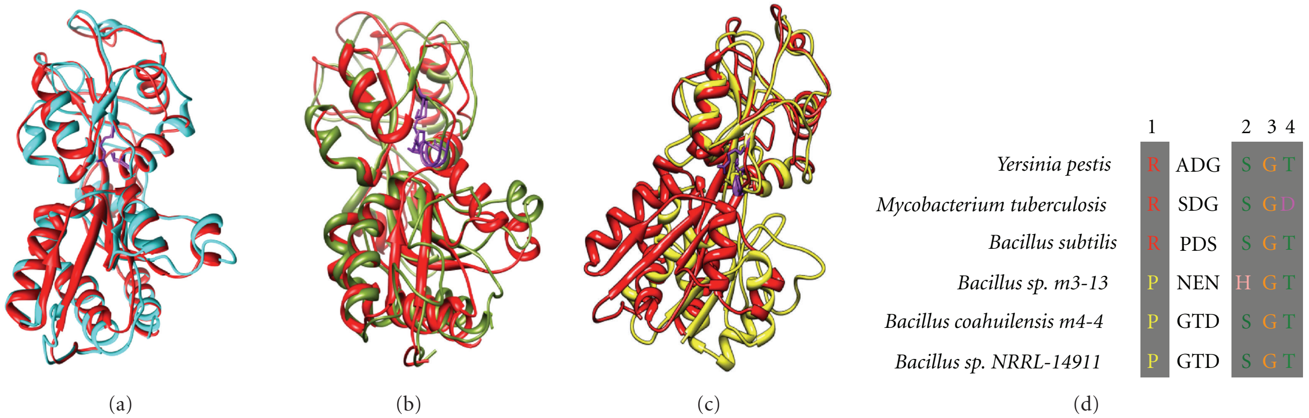

(a)
(b)
(c)
(d)
1 2 3 4
Yersinia pestis R ADG S G T
Mycobacterium tuberculosis R SDG S G D
Bacillus subtilis R PDS S G T
Bacillus sp. m3-13 P NEN H G T
Bacillus coahuilensis m4-4 P GTD S G T
Bacillus sp. NRRL-14911 P GTD S G T

Figure 5: Comparison of 3D structures with TOPOFIT of PstS from *B. subtilis* (red), with (a) *B. sp. NRRL-14911* (cyan; r.m.s.d. = 1.02, z-score = 39.89), (b) *Bacillus sp. m3-13* (green; r.m.s.d. = 1.53, z-score = 20.24) and (c) *B. coahuilensis m4-4* (yellow; r.m.s.d. = 1.49, z-score = 9.65). Residues involved in phosphate binding are highlighted in purple. (d) shows an alignment of the active site of PstS and the different amino acids involved in phosphate binding are highlighted in gray.

3 billion years ago ([37]; this study). Even though we observed a constant operon architecture, we also observed two taxa with both types of operon (*S. thermophilum* and *S. pneumoniae*), one early divergent and the other more derived (Figure 3), however, the protein sequence of either of them is very divergent from the other taxa.

Even if the general architecture of the operon suggests less recombination than in the *Cyanobacteria* lineages, in *Firmicutes* several other taxa had more than one copy of the *pstS* gene or presented an incomplete extra copy of the operon. A phylogenetic analysis of PstS (see Figure S1 in Supplementary Material available online at doi:10.406/2011/781642) suggests that these extra copies of the *pstS* gene were acquired later by HGT rather than acquired by duplications, since different copies of the gene belonging to the same organism are found in different places in the phylogeny. Even though the high sequence variation present in *pstS* at both nucleotide and protein levels make it hard to obtain a well-supported phylogenetic reconstruction. From our analysis it is evident that the *pstS* gene is evolving at a faster rate than the rest of the genes of the operon, where the purifying selection maintains the folding of the protein instead of the amino acid sequence [19]. Since we found no significant positive selection for *pstS*, the high level of sequence divergence could be due to the accumulation of repeated mutations after the ancient split between the *cereus*-like and *subtilis*-like operon [50].

Interestingly, despite the marked sequence divergence of the genes of the *pst* operon of firmicutes and particularly the *pstS* gene (Figure 4), the 3D structures of PstS from the *subtilis*-like and *cereus*-like operons were surprisingly similar (Figures 5(a)–5(c)). However, it is still to be determined how the changes of particular amino acids in the phosphate binding site (Figure 5(d)) would affect the formation of the hydrogen bonds necessary for phosphate uptake [51]. In particular, it is possible that the presence of a proline (P) instead of serine in the active site of the *cereus*-like PstS (Figure 5(d)) could have some relevance in the discrimination between the mono- and dibasic forms of phosphate, since this amino acid only acts as a hydrogen bond acceptor but not as a donor [46]. The difference of affinity to phosphate and the potential selectivity for either of the phosphate species should be investigated experimentally, as the protein backbone is also involved in the hydrogen bond formation and not only the side chains of the residues [46, 51]. The amino acid changes in the active site of the PstS protein can have an effect on the efficiency of the acceptor under different pH conditions [46, 51], and maybe some of those substitutions could be related to habitat. Nevertheless, this idea is not sustained within the *cereus*-like clade, were it can be observed that the lineage with *B. sp m3-13* (Figure 2), the *Staphylococcus* group and the clade with *Anoxyblacillus flavithermus* as well as two species of *Geobacillus* have a histidine instead of a serine in the second residue of the active site (Figure 5(d)), and all those taxa actually live in environments with a wide range of pH conditions, as is the case in the rest of *Firmicutes*.

It has been previously noted that protein structure is fairly conserved in nature, and that proteins with only 8% similarity at sequence level can have a much higher similarity in their structural features [52, 53]. In our case, some of the PstS proteins had a sequence similarity as low as 17% when comparing those from the *subtilis*- or *cereus*-like operons, yet their structure was fairly conserved (Figures 5(a)–5(c)). This could be due to natural selection acting on protein structure, thus allowing for changes in amino acids that would not alter the basic features of the protein [54]. The fact that these similar protein structures occur on lineages that have such deep divergences such as *Cyanobacteria* and *Firmicutes* (ca. 3 billion years ago), favors the idea of parallel evolution from a common ancestor [37, 54, 55]. This very ancient divergence produced one *pst* operon mostly found in anaerobic *Firmicutes* and some pathogenic groups (*Listeria*), while a *pst* operon similar to that of *Cyanobacteria* is found mostly in oligotrophic *Firmicutes* and in some other pathogenic groups (*Staphylococcus*), with various cases of ancient HGT between either group (i.e., *B. halodurans*, *B. clausii*, and *D. hafniense*).

Therefore, we should reconsider the environmental constrain hypothesis: *Bacillus sp.* CH108 from the Churince water system in the present shares similar environmental conditions to other *Bacillus* from CCB, but has a *subtilis*-like *pst* operon instead of the *cereus*-like operon that is common to other CCB species. The sister species of *B. sp.* CH108, *B. marisflavi* from the Yellow Sea of Korea [56], also has a *subtilis*-like operon. This suggests that acquisition of the operon predates the divergence of the two taxa (∼92 my ago; Figure 3), which in turn is older than the last time CCB was connected to the ocean, ca. 45 my ago. This may leads to the idea that this particular arrangement is not specific to the actual oligotrophic conditions in Cuatro Cienegas [41, 57] but it is an adaptation to an ancient sea.

5. Conclusions

The *pst* operon in *Firmicutes* showed a very high sequence divergence that is not correlated to either phylogenetic relationships among taxa, the type of habitat, or the phosphorus availability where these organisms currently live. Thus, it is likely that the current distribution the *pst* operon was determined by a very early divergence and repeated events of HGT of the phosphate transporter genes followed by parallel evolution that lead to similar 3D structures. Unlike what was observed in *Cyanobacteria*, most *Firmicutes* only have one or a couple of copies of the PstS protein, so it is crucial for phosphate uptake that both function and affinity are conserved in the substrate binding protein.

Acknowledgments

This project was funded by CONACYT-SEP Grant no. 57507 and CONACyT-Semarnat 2006-C01-23459 awarded to V. Souza and a CINVESTAV Multidisciplinary awarded to G. Olmedo. V. Souza and L.E. Eguiarte worked in this paper during a sabbatical leave supported by DGAPA and UC-Mexus, respectively. Thanks are due to L. D. Alcaraz, E. Rebollar, F. García-Oliva, D. Ortega-Del Vecchyo, I. Hernandez, V. Lopez, G. Moreno-Hagelsieb, B.S. Gaut, M. Tenaillon, O. Tenaillon, A. Vázquez-Lobo, and D. Piñero for all their insights and comments.

References

[1] S. G. Tetu, B. Brahamsha, D. A. Johnson et al., “Microarray analysis of phosphate regulation in the marine cyanobacterium *Synechococcus sp. WH8102*,” *ISME Journal*, vol. 3, no. 7, pp. 835–849, 2009.

[2] P. G. Falkowski, T. Fenchel, and E. F. Delong, “The microbial engines that drive earth’s biogeochemical cycles,” *Science*, vol. 320, no. 5879, pp. 1034–1039, 2008.

[3] D. Papineau, “Global biogeochemical changes at both ends of the Proterozoic: insights from phosphorites,” *Astrobiology*, vol. 10, no. 2, pp. 165–181, 2010.

[4] J. J. Elser, J. Watts, J. H. Schampel, and J. Farmer, "Early Cambrian food webs on a trophic knife-edge? A hypothesis and preliminary data from a modern stromatolite-based ecosystem.," *Ecology Letters*, vol. 9, no. 3, pp. 295–303, 2006.

[5] J. J. Elser and A. Hamilton, "Stoichiometry and the new biology: the future is now," *PLoS Biology*, vol. 5, no. 7, article e181, 2007.

[6] M. V. Zubkov, I. Mary, E. M. S. Woodward et al., "Microbial control of phosphate in the nutrient-depleted North Atlantic subtropical gyre," *Environmental Microbiology*, vol. 9, no. 8, pp. 2079–2089, 2007.

[7] D. B. Rusch, A. L. Halpern, G. Sutton et al., "The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific," *PLoS Biology*, vol. 5, no. 3, article e77, 2007.

[8] M. M. Adams, M. R. Gómez-García, A. R. Grossman, and D. Bhaya, "Phosphorus deprivation responses and phosphonate utilization in a thermophilic *Synechococcus* sp. from microbial mats," *Journal of Bacteriology*, vol. 190, no. 24, pp. 8171–8184, 2008.

[9] L. D. Alcaraz, G. Olmedo, G. Bonilla et al., "The genome of *Bacillus coahuilensis* reveals adaptations essential for survival in the relic of an ancient marine environment," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, no. 15, pp. 5803–5808, 2008.

[10] Y. Qi, Y. Kobayashi, and F. M. Hulett, "The *pst* operon of *Bacillus subtilis* has a phosphate-regulated promoter and is involved in phosphate transport but not in regulation of the Pho regulon," *Journal of Bacteriology*, vol. 179, no. 8, pp. 2534–2539, 1997.

[11] A. C. Martiny, M. L. Coleman, and S. W. Chisholm, "Phosphate acquisition genes in *Prochlorococcus* ecotypes: evidence for genome-wide adaptation," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, no. 33, pp. 12552–12557, 2006.

[12] M. Aguena and B. Spira, "Transcriptional processing of the *pst* operon of *Escherichia coli*," *Current Microbiology*, vol. 58, no. 3, pp. 264–267, 2009.

[13] A. C. Martiny, A. P. K. Tai, D. Veneziano, F. Primeau, and S. W. Chisholm, "Taxonomic resolution, ecotypes and the biogeography of *Prochlorococcus*," *Environmental Microbiology*, vol. 11, no. 4, pp. 823–832, 2009.

[14] M. Sebastian and J. W. Ammerman, "The alkaline phosphatase PhoX is more widely distributed in marine bacteria than the classical PhoA," *ISME Journal*, vol. 3, no. 5, pp. 563–572, 2009.

[15] M. Aguena, E. Yagil, and B. Spira, "Transcriptional analysis of the *pst* operon of *Escherichia coli*," *Molecular Genetics and Genomics*, vol. 268, no. 4, pp. 518–524, 2002.

[16] N. E. E. Allenby, N. O'Connor, Z. Prágai et al., "Post-transcriptional regulation of the *Bacillus subtilis pst* operon encoding a phosphate-specific ABC transporter," *Microbiology*, vol. 150, no. 8, pp. 2619–2628, 2004.

[17] R. J. Fischer, S. Oehmcke, U. Meyer et al., "Transcription of the pst operon of *Clostridium acetobutylicum* is dependent on phosphate concentration and pH," *Journal of Bacteriology*, vol. 188, no. 15, pp. 5469–5478, 2006.

[18] K. Tomii and M. Kanehisa, "A comparative analysis of ABC transporters in complete microbial genomes," *Genome Research*, vol. 8, no. 10, pp. 1048–1059, 1998.

[19] A. L. Davidson, E. Dassa, C. Orelle, and J. Chen, "Structure, function, and evolution of bacterial ATP-binding cassette systems," *Microbiology and Molecular Biology Reviews*, vol. 72, no. 2, pp. 317–364, 2008.

[20] A. C. Martiny, Y. Huang, and W. Li, "Occurrence of phosphate acquisition genes in *Prochlorococcus* cells from different ocean regions," *Environmental Microbiology*, vol. 11, no. 6, pp. 1340–1347, 2009.

[21] R. Feingersch, M. T. Suzuki, M. Shmoish et al., "Microbial community genomics in eastern Mediterranean Sea surface waters," *ISME Journal*, vol. 4, no. 1, pp. 78–87, 2010.

[22] B. A. S. Van Mooy, H. F. Fredricks, B. E. Pedler et al., "Phytoplankton in the ocean use non-phosphorus lipids in response to phosphorus scarcity," *Nature*, vol. 458, no. 7234, pp. 69–72, 2009.

[23] J. J. Elser, J. H. Schampel, F. Garcia-Pichel et al., "Effects of phosphorus enrichment and grazing snails on modern stromatolitic microbial communities," *Freshwater Biology*, vol. 50, no. 11, pp. 1808–1825, 2005.

[24] V. Souza, L. E. Eguiarte, J. Siefert, and J. J. Elser, "Microbial endemism: does phosphorus limitation enhance speciation?" *Nature Reviews Microbiology*, vol. 6, no. 7, pp. 559–564, 2008.

[25] V. Souza, L. Espinosa-Asuar, A. E. Escalante et al., "An endangered oasis of aquatic microbial biodiversity in the Chihuahuan desert," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, no. 17, pp. 6565–6570, 2006.

[26] C. Desnues, B. Rodriguez-Brito, S. Rayhawk et al., "Biodiversity and biogeography of phages in modern stromatolites and thrombolites," *Nature*, vol. 452, no. 7185, pp. 340–343, 2008.

[27] L. D. Alcaraz, G. Moreno-Hagelsieb, L. E. Eguiarte, V. Souza, L. Herrera-Estrella, and G. Olmedo, "Understanding the evolutionary relationships and major traits of *Bacillus* through comparative genomics," *BMC Genomics*, vol. 11, no. 1, article no. 332, 2010.

[28] M. Breitbart, A. Hoare, A. Nitti et al., "Metagenomic and stable isotopic analyses of modern freshwater microbialites in Cuatro Ciénegas, Mexico," *Environmental Microbiology*, vol. 11, no. 1, pp. 16–34, 2009.

[29] B. M. Winsborough, E. Theriot, and D. B. Czarnecki, *Diatoms on a Continental Island: Lazarus Species, Marine Disjuncts and other Endemic Diatoms of the Cuatro Ciénegas basin, Coahuila, México*, University of Texas, Austin Tex, USA, 2008.

[30] W. L. Minkley, "Cuatro Cienegas fishes: research reviewd and a local test of diversity versus habitat size," *Journal of the Arizona-Nevada Academy of Science*, vol. 19, pp. 13–21, 1984.

[31] E. W. Carson and T. E. Dowling, "Influence of hydrogeographic history and hybridization on the distribution of genetic variation in the pupfishes *Cyprinodon atrorus* and *C. bifasciatus*," *Molecular Ecology*, vol. 15, no. 3, pp. 667–679, 2006.

[32] S. G. Johnson, "Age, phylogeography and population structure of the microendemic banded spring snail, *Mexipyrgus churinceanus*," *Molecular Ecology*, vol. 14, no. 8, pp. 2299–2311, 2005.

[33] M. Tobler and E. W. Carson, "Environmental variation, hybridization, and phenotypic diversification in Cuatro Ciénegas pupfishes," *Journal of Evolutionary Biology*, vol. 23, no. 7, pp. 1475–1489, 2010.

[34] R. Cerritos, P. Vinuesa, L. E. Eguiarte et al., "*Bacillus coahuilensis* sp. nov., a moderately halophilic species from a desiccation lagoon in the Cuatro Ciénegas Valley in Coahuila, Mexico," *International Journal of Systematic and Evolutionary Microbiology*, vol. 58, no. 4, pp. 919–923, 2008.

[35] R. Cerritos, L. E. Eguiarte, M. Avitia et al., "Diversity of culturable thermo-resistant aquatic bacteria along an environmental gradient in Cuatro Ciénegas, Coahuila, Mexico," *Antonie Van Leeuwenhoek*, vol. 99, no. 2, pp. 303–318, 2010.

[36] A. E. Escalante, L. E. Eguiarte, L. Espinosa-Asuar, L. J. Forney, A. M. Noguez, and V. Souza Saldivar, "Diversity of aquatic prokaryotic communities in the Cuatro Cienegas basin," *FEMS Microbiology Ecology*, vol. 65, no. 1, pp. 50–60, 2008.

[37] F. U. Battistuzzi, A. Feijao, and S. B. Hedges, "A genomic timescale of prokaryote evolution: insights into the origin of methanogenesis, phototrophy, and the colonization of land," *BMC Evolutionary Biology*, vol. 4, article no. 44, 2004.

[38] H. Maughan, "Rates of molecular evolution in bacteria are relatively constant despite spore dormancy," *Evolution*, vol. 61, no. 2, pp. 280–288, 2007.

[39] F. Kunst, N. Ogasawara, I. Moszer et al., "The complete genome sequence of the gram-positive bacterium *Bacillus subtilis*," *Nature*, vol. 390, no. 6657, pp. 249–256, 1997.

[40] M. J. Sanderson, "Estimating absolute rates of molecular evolution and divergence times: a penalized likelihood approach," *Molecular Biology and Evolution*, vol. 19, no. 1, pp. 101–109, 2002.

[41] I. Ferrusquía-Villafranca, "Geología de México: una sinopsis," in *Diversidad biológica de México: orígenes y distribución*, T. P. Ramamoorthy et al., Ed., Instituto de Biología UNAM, México D.F., 1998.

[42] R. C. Edgar, "MUSCLE: a multiple sequence alignment method with reduced time and space complexity," *BMC Bioinformatics*, vol. 5, article no. 113, 2004.

[43] A. Stamatakis, P. Hoover, and J. Rougemont, "A rapid bootstrap algorithm for the RAxML web servers," *Systematic Biology*, vol. 57, no. 5, pp. 758–771, 2008.

[44] F. Abascal, R. Zardoya, and D. Posada, "ProtTest: selection of best-fit models of protein evolution," *Bioinformatics*, vol. 21, no. 9, pp. 2104–2105, 2005.

[45] T. L. Bailey and C. Elkan, "Fitting a mixture model by expectation maximization to discover motifs in biopolymers," in *Proceedings of the 2nd International Conference on Intelligent Systems for Molecular Biology*, vol. 2, pp. 28–36, 1994.

[46] M. Tanabe, O. Mirza, T. Bertrand et al., "Structures of OppA and PstS from *Yersinia pestis* indicate variability of interactions with transmembrane domains," *Acta Crystallographica Section D*, vol. 63, no. 11, pp. 1185–1193, 2007.

[47] N. Eswar, B. Webb, M. A. Marti-Renom et al., "Comparative protein structure modeling using Modeller," *Current Protocols in Bioinformatics*, vol. 5, pp. 5.6.1–5.6.30, 2006.

[48] V. A. Ilyin, A. Abyzov, and C. M. Leslin, "Structural alignment of proteins by a novel TOPOFIT method, as a superimposition of common volumes at a topomax point," *Protein Science*, vol. 13, no. 7, pp. 1865–1874, 2004.

[49] K. Tamura, J. Dudley, M. Nei, and S. Kumar, "MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0," *Molecular Biology and Evolution*, vol. 24, no. 8, pp. 1596–1599, 2007.

[50] A. L. Hughes, "Looking for Darwin in all the wrong places: the misguided quest for positive selection at the nucleotide sequence level," *Heredity*, vol. 99, no. 4, pp. 364–373, 2007.

[51] H. Luecke and F. A. Quiocho, "High specificity of a phosphate transport protein determined by hydrogen bonds," *Nature*, vol. 347, no. 6291, pp. 402–406, 1990.

[52] E. V. Koonin, Y. I. Wolf, and G. P. Karev, "The structure of the protein universe and genome evolution," *Nature*, vol. 420, no. 6912, pp. 218–223, 2002.

[53] R. A. Goldstein, "The structure of protein evolution and the evolution of protein structure," *Current Opinion in Structural Biology*, vol. 18, no. 2, pp. 170–177, 2008.

[54] A. Sánchez-Flores, E. Pérez-Rueda, and L. Segovia, "Protein homology detection and fold inference through multiple alignment entropy profiles," *Proteins: Structure, Function and Genetics*, vol. 70, no. 1, pp. 248–256, 2008.

[55] R. Woods, D. Schneider, C. L. Winkworth, M. A. Riley, and R. E. Lenski, "Tests of parallel molecular evolution in a long-term experiment with *Escherichia coli*," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, no. 24, pp. 9107–9112, 2006.

[56] J. H. Yoon, I. G. Kim, K. H. Kang, T. K. Oh, and Y. H. Park, "*Bacillus marisflavi* sp. nov. and *Bacillus aquimaris* sp. nov., isolated from sea water of a tidal flat of the Yellow Sea in Korea," *International Journal of Systematic and Evolutionary Microbiology*, vol. 53, no. 5, pp. 1297–1303, 2003.

[57] F. Vega, T. Nyborg, M. Perrilliat, M. Montellanos-Ballesteros, S. R. S. Cevallos-Ferriz, and S. A. Quiroz-Barroso, *Studies on Mexican Paleontology*, Springer, Dordrecht, The Nederlands, 2006.

SAGE-Hindawi Access to Research
International Journal of Evolutionary Biology
Volume 2011, Article ID 970768, 11 pages
doi:10.4061/2011/970768

Research Article

Ectopic Gene Conversions in the Genome of Ten Hemiascomycete Yeast Species

Robert T. Morris and Guy Drouin

Département de Biologie et Centre de Recherche Avancée en Génomique Environnementale, Université d'Ottawa, 30 Marie Curie, Ottawa, ON, Canada K1N 6N5

Correspondence should be addressed to Guy Drouin, gdrouin@science.uottawa.ca

Received 21 July 2010; Revised 18 September 2010; Accepted 15 October 2010

Academic Editor: Hiromi Nishida

We characterized ectopic gene conversions in the genome of ten hemiascomycete yeast species. Of the ten species, three diverged prior to the whole genome duplication (WGD) event present in the yeast lineage and seven diverged after it. We analyzed gene conversions from three separate datasets: paralogs from the three pre-WGD species, paralogs from the seven post-WGD species, and common ohnologs from the seven post-WGD species. Gene conversions have similar lengths and frequency and occur between sequences having similar degrees of divergence, in paralogs from pre- and post-WGD species. However, the sequences of ohnologs are both more divergent and less frequently converted than those of paralogs. This likely reflects the fact that ohnologs are more often found on different chromosomes and are evolving under stronger selective pressures than paralogs. Our results also show that ectopic gene conversions tend to occur more frequently between closely linked genes. They also suggest that the mechanisms responsible for the loss of introns in *S. cerevisiae* are probably also involved in the gene $3'$-end gene conversion bias observed between the paralogs of this species.

1. Introduction

The repair of double strand DNA breaks is a critical biological process which maintains genome stability. The primary process whereby double-strand DNA breaks are repaired is via homologous recombination; this process requires the use of a repair template gene which provides a copy of the missing information caused by the double-strand DNA breaks. The repair template can either be an allele (allelic recombination) or a paralog (ectopic recombination). An end product of the homologous recombination pathway is the replacement of the broken part of the damaged gene by a homologous portion of the repair template gene. The damaged gene is therefore converted by the template gene (reviewed in [1]).

The factors affecting, and the characteristics of, ectopic and allelic gene conversions have been the focus of many studies, and sequence similarity has been shown to have a profound effect on gene conversion propensity between paralogs. In *Escherichia coli,* a 2%–4% decrease in sequence similarity between a damaged gene and its repair template can cause a 10- to 40-fold decrease in recombination frequency [2, 3]. Similarly, in *Saccharomyces cerevisiae*, larger gene conversions are limited to more similar sequences [4]. Chromosomally linked genes are converted more frequently than dispersed genes in Drosophila and humans [5, 6]. In *S. cerevisiae*, increasing distance between paralogs located on the same chromosome tends to decrease their conversion frequency [4, 7, 8]. In some genomes, different regions of genes are converted at different rates. For example, in *S. cerevisiae*, genes conversions between dispersed paralogs are more frequent at their $3'$-ends [4]. This $3'$-bias is likely the result of gene conversion with incomplete cDNA molecules [9].

The availability of ten hemiascomycete genomes provides the opportunity to study ectopic gene conversions within a clade with as much sequence divergence as the entire Chordate phylum [10]. The evolution of several hemiascomycetes species was affected by a whole genome duplication event (WGD) which occurred some 150 millions

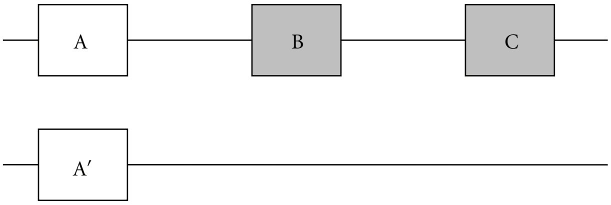


Figure 1: Schematic representation of ohnologs and paralogs. Genes A and A′ represent ohnologs created by a genome duplication. These genes are therefore located on different chromosomes. Genes B and C represent paralogs created by tandem duplications of gene A. These genes are therefore on the same chromosome as gene A.

years ago (MYA; [11–14]). The genomes of *Kluyveromyces lactis*, *Debaryomyces hansenii,* and *Yarrowia lipolytica* all diverged before the whole genome duplication event that occurred in the yeast lineage (pre-WGD species; [10]). The *S. cerevisiae*, *Saccharomyces paradoxus*, *Saccharomyces mikatae*, *Saccharomyces bayanus*, *Saccharomyces kudriavzevii*, *Saccharomyces castellii,* and *Candida glabrata* genomes all diverged after this whole genome duplication event (post-WGD species; [15–17]).

The advantage of separating these genomes into two groups is that we are able to perform two comparisons. The first compares the characteristics of ectopically converted ohnologs and paralogs between the post-WGD species. The post-WGD ohnologs are composed of the duplicated gene pairs that resulted from the whole genome duplication [11, 18]. The post-WGD paralogs data set is composed of the genes from multigene families containing at least three members in the genome of the seven post-WGD species but excluding all ohnologs (Figure 1). The second comparison involves the contrast of the characteristics of ectopically converted paralogs between pre- and post-WGD species. The pre-WGD paralogs data set is composed of the genes from multigene families containing at least three members in the genome of the three pre-WGD species.

The previous studies have shown that the reason why many ohnologs are still found in yeast genomes is because they provide a selective advantage [19, 20]. Ohnologs are maintained by selection either because they carry out a subset of the functions that were previously assumed by their preduplication ancestor (subfunctionalization), assume new functions (neofunctionalization), or provide increased gene product dosage. We therefore expect that most ectopic gene conversions between ohnolog genes will be deleterious and removed by selection. If so, ectopic gene conversions between ohnologs should be less frequent than those between paralogs. In addition, based on the previous studies, we expect that gene conversion frequency should decrease as the distance between related genes increases (and be least frequent for genes situated on different chromosomes), that the length of gene conversion tracts should be positively correlated with sequence similarity and that converted regions should be more frequent at the 3′-end of genes [4].

2. Materials and Methods

2.1. Genome Sequences. The *S. cerevisiae*, *S. paradoxus*, *S. mikatae*, *S. bayanus*, *S. kudriavzevii*, and *S. castellii* genome sequences were retrieved from the Saccharomyces Genome Database (SGD; ftp://genome-ftp.stanford.edu/pub/yeast/sequence/). The *C. glabrata*, *K. lactis*, *D. hansenii*, and *Y. lipolytica* genome sequences and distance files (*.ptt files) were retrieved from the NCBI ftp website (ftp://ftp.ncbi.nih.gov/).

2.2. Gene Family Data Sets. We used three different data sets of protein coding genes. To retrieve the post-WGD ohnologs from the seven post-WGD species, we used the 551 *S. cerevisiae* duplicated gene pairs (1102 ohnologs) identified by Byrne and Wolfe [21] as queries. Sequences from *C. glabrata* and *S. castellii* were retrieved using the Yeast Gene Order Browser (http://wolfe.gen.tcd.ie/ygob/), and those from the other 4 species were retrieved from the Saccharomyces Genome Database (ftp://genome-ftp.stanford.edu/pub/yeast/sequence/fungal_genomes/Multiple_species_align/other/fungalAlignCorrespondance.txt). Our data set of ohnologs in post-WGD species is therefore only composed of the ohnologs pairs also found in *S. cerevisiae*. We used this subset of ohnologs because the efficient detection of gene conversion events using the GENECONV method requires that at least three sequences be available [4]. To detect gene conversions in ohnologs, we therefore needed ohnologs from at least two species and we used the ohnologs of *S. cerevisiae* to retrieve ohnologs pairs from the other 6 post-WGD species. Retrieving common ohnologs also allowed us to study gene conversions between similar genes in seven different genomes.

The post-WGD paralog data set was constructed using the BLASTCLUST program available at the NCBI FTP site. Gene families were defined as being composed of sequences having at least 60% amino acid identity over at least 50% of their length. If genes previously identified as ohnologs were grouped into paralog multigene families, then these genes were removed from the family to ensure that there was no redundancy between the ohnolog and paralog data sets (see Figure 1). The pre-WGD paralog data set was also constructed using the BLASTCLUST program, and gene families were also defined as being composed of sequences having at least 60% amino acid identity over at least 50% of their length.

2.3. Sequence Alignments and Gene Conversion Detection. ClustalW was used to align the protein sequences of multigene families' members [22]. DNA sequences were then fitted to the protein alignments using a PERL script.

Gene conversions were detected using the GENECONV method [23]. Redundant gene conversions within a multigene family were detected by examining the phylogenetic tree of each family and removed from the analysis [4]. If the same gene conversion was detected at the same location in the multigene family alignment in closely related descendents of a common ancestor then the most parsimonious

Table 1: Number of ohnologs and paralogs in the pre- and post-WGD genomes.

Genome	Number of ohnolog families	Number of paralog families
Post-WGD		
S. cerevisiae	551 (2)	30 (3–40)
S. paradoxus	436 (2)	80 (3–68)
S. mikatae	412 (2)	86 (3–37)
S. kudriavzevii	226 (2)	13 (3–20)
S. bayanus	462 (2)	75 (3–23)
C. glabrata	300 (2)	16 (3–7)
S. castellii	398 (2)	17 (3–10)
Pre-WGD		
K. lactis	n.a.	15 (3–9)
D. hansenii	n.a.	43 (3–9)
Y. lipolytica	n.a.	60 (3–26)

Notes. The range of multigene family sizes is provided in brackets. n.a.: not applicable.

explanation is that the conversion event occurred within the common ancestor, therefore only one of the conversions detected in the set of descendents was retained for further analysis. To control for false positives, gene conversions between sequences having less than 80% maximum flanking similarity were removed from the analysis [24].

2.4. Gene Conversion Characteristics. The gene conversion frequency for each species was calculated using two different methods. The first method calculates the conversion frequency as the ratio of the number of conversions divided by the total number of gene comparisons between multigene family members. The second method calculates the frequency as the ratio of the number of gene conversions divided by the total number of multigene family members. Intra- and interchromosomal gene conversion frequencies were calculated for the *S. cerevisiae*, *C. glabrata* ohnolog and paralog multigene families. In addition intra- and interchromosomal conversion frequencies were calculated for the paralog multigene families of *K. lactis*, *D. hansenii*, and *Y. lipolytica* genomes. These frequencies are calculated as the ratio of intra- (or inter-) chromosomal conversions divided by the total number of intra- (or inter-) chromosomal gene comparisons. The gene conversion length was obtained from the GENECONV output. The maximum similarity for the flanking 100 nucleotides was calculated for each converted gene pair using an in-house PERL script. The locations of the converted regions were calculated as the correlation between the positions of each conversion with respect to the length of the converted genes. A positive correlation indicates a bias towards the 3′-end of genes, and a negative correlation indicates a bias towards the 5′-end of genes. The distance between converted genes was calculated only for conversions detected within *S. cerevisiae*, *C. glabrata*, *K. lactis, D. hansenii*, and *Y. lipolytica* because position data for the other five species was not available. Data tabulation and analysis was done using Microsoft Excel (Microsoft, Redmond, WA, USA) and S-plus v 7.0 (Insightful, Seattle, WA, USA). The G-Power program was used to calculate the power of the ANOVA tests [25]. Power calculations for correlation tests were done using an online application (http://calculators.stat.ucla.edu/powercalc/correlation/) and SAS 9.1.3 (SAS Institute Inc., Cary, NC, USA).

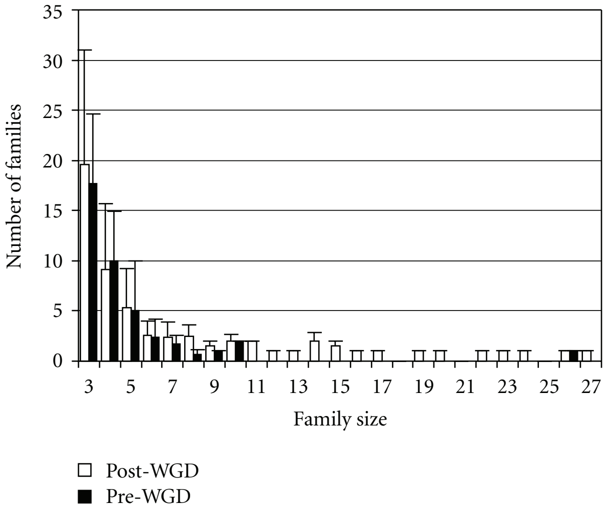


Figure 2: The distribution of the average number of paralog gene families (mean ± S.D.) within the seven postduplication genomes and three preduplication genomes is shown. Five outlier families including two families of size 63 and 68 from *S. paradoxus*, two families of size 32 and 40 from *S. cerevisiae*, and a single family of 38 genes from *S. mikatae* are not shown in the figure to improve the visual clarity of the data.

2.5. Numbers of Substitutions per Site and Gene Ontology. The number of nonsynonymous substitutions per nonsynonymous site (Ka) and synonymous substitutions per synonymous site (Ks) and their ratio (Ka/Ks) were calculated for the protein coding regions (excluding the converted regions) of each pair of converted genes using the YN00 program from the PAML software [26, 27].

The processes in which the *S. cerevisiae* ohnologs and paralogs are involved were analyzed using the gene ontology annotations of the Saccharomyces Genome Database at http://db.yeastgenome.org/cgi-bin/GO/goTermFinder.pl [28].

3. Results

3.1. Ohnolog and Paralog Multigene Families. Ohnolog and paralog multigene families were analyzed to determine whether the number and size of these two types of families were different in different yeast genomes (Table 1, Figure 2). The genomes of the six post-WGD species from which we retrieved ohnolog pairs using the *S. cerevisiae* ohnologs contain an average of 372 ± 90 ohnolog pairs. The number of ohnolog pairs found in each of these six different genomes is not significantly different from this average when using a

Table 2: Percentage of gene comparisons between multigene family members located on the same chromosome.

Genome	Ohnologs	Paralogs
Post-WGD		
S. cerevisiae	4.0% (22/551)	8.4% (163/1930)
C. glabrata	5.0% (15/300)	38.6% (29/75)
Pre-WGD		
K. lactis	n.a.	21% (26/124)
D. hansenii	n.a.	31% (86/270)
Y. lipolytica	n.a.	18% (158/884)

Notes. The ratios in brackets are the number of gene comparisons between genes found on the same chromosome divided by the total number of gene comparisons. n.a.: not applicable.

Bonferroni-corrected α-value of 0.0083 (Wilcoxon rank sum test; [29]).

For post-WGD paralogs, only the *S. mikatae* genome has significantly more paralog families than average (45.28 ± 33.36; Wilcoxon rank sum test, $P = 0.009$) and only the *S. kudriavzevii* genome has significantly fewer paralog families than average ($P = 0.009$). The mean size of the paralog families (5.7 ± 5.2 genes/family) is similar in all post-WGD genomes except that of *C. glabrata* which has significantly smaller paralog families than average (3.3±0.99 genes/family, Wilcoxon rank sum test, $P = 0.003$).

For the pre-WGD paralogs, the numbers of paralog families in the three pre-WGD genomes are not significantly different from the population mean (39.33±22.72; Wilcoxon rank sum test, $P \geq 0.27$). The mean size of all paralog families in these three genomes (4.41 ± 2.59 paralogs per family) is similar to the mean family size of each pre-WGD genome (Wilcoxon rank sum test, $P \geq 0.09$). Finally, there is no statistical difference between the number (Wilcoxon rank sum test, $P = 0.83$) and the mean size of paralog families (Wilcoxon rank sum test, $P = 0.17$) or between pre- and post-WGD species.

3.2. Organization of Gene Families. The organization of the multigene families can be measured as the proportion of multigene family members located on the same chromosome. Since most paralogs originate from unequal crossover events, they are expected to be most often found on the same chromosome. In contrast, since ohnologs are remnants of ancient genome duplication events, they are expected to be most often found on different chromosomes. The higher percentage of paralogs found on the same chromosome is therefore consistent with the likely mode of origin of these two types of duplicated genes (Table 2). The percentage of paralogs found on the same chromosome is also similar between pre- and post-WGD genomes (Table 2).

3.3. Gene Conversion Frequency and Distance between Converted Genes. In post-WGD genomes, intrachromosomal gene conversions tend to occur more frequently than interchromosomal conversions. In the paralog families of *S. cerevisiae* and *C. glabrata*, genes located on the same chromosome are converted 2 to 10 times more frequently than genes found on different chromosomes (Table 3). Similarly, in the ohnolog families of *S. cerevisiae*, genes located on the same chromosome are converted 4 times more frequently than genes found on different chromosomes (Table 3). In contrast, there is an almost complete absence of gene conversions between the ohnologs found within the *C. glabrata* genome (Table 3).

In pre-WGD genomes, the paralogs found on the same chromosomes of *K. lactis* and *Y. lipolytica* are not converted more frequently than paralogs found on different chromosomes but the *D. hansenii* paralogs found on the same chromosomes are converted roughly 3 times more frequently than those found on different chromosomes (Table 3).

The mean number (±S.D.) of conversions detected within the paralog gene families of the pre- (38 ± 33) and post-WGD (30 ± 16) genomes is not statistically different (t-test, $P = 0.67$; Table 4). Although the ohnolog families of post-WGD genomes contain only an average of 7 ± 5 conversions, this number is also not significantly different from the average number of conversions found in post-WGD paralog families (t-test, $P = 0.06$).

When considering gene conversion frequencies with respect to the total number of comparisons, gene conversions of post-WGD species are either equally frequent in paralog and ohnolog families (in the *S. paradoxus*, *S. mikatae,* and *S. bayanus* genomes) or significantly more frequent in paralog than in ohnolog families in the four other post-WGD families (t-test, $P = 0.046$; Table 4).

When considering gene conversion frequencies with respect to the total number of multigene family members, the mean conversion frequency for paralogs (19.03 ± 16.29%) is significantly larger than the frequency for ohnologs (0.74 ± 0.46; Wilcoxon two sample test, $P = 0.0006$).

We believe that using gene conversion frequencies with respect to the total number of multigene family members is more appropriate to compare gene frequencies between ohnologs and paralogs because it better reflects the much larger number of conversions found in paralogs when compared to ohnologs. For example, in the case of *S. cerevisiae* with 13 conversions between ohnologs and 110 conversion between paralogs (Table 4), the conversion frequency for ohnologs is 2.35% (13/551) and 5.71% for paralogs (110/1930) when frequencies are calculated with respect to the total number of comparisons. However, these frequencies do not take into account the fact that 1102 ohnolog sequences were compared (551 pairs) whereas only 212 paralog sequences (i.e., less than the fifth of the number of ohnolog sequences) were compared (for a total of 1930 pairwise comparisons) to obtain the 5.71% frequency of paralogs. In contrast, if one compares the frequencies calculated with respect with the number of genes, the frequency of conversions is 1.17% (13/1102) for ohnologs and 51.40% for paralogs (110/212). The large difference between the two ways of calculating frequencies is due to the fact that frequencies calculated with respect to the total number of comparisons have a much larger denominator which biases the comparisons between ohnologs and paralogs. For example, for a family with 10 paralogous sequences, the number of pairwise comparisons

Table 3: Intra- and interchromosomal gene conversion frequencies for pre- and post-WGD genomes.

Genome	Ohnologs		Paralogs	
Post-WGD	Intrachromosomal frequency	Interchromosomal frequency	Intrachromosomal frequency	Interchromosomal frequency
S. cerevisiae	9.1% (2/22)	2.1% (11/529)	9.2% (15/163)	5.4% (95/1767)
C. glabrata	0% (0/15)	0.007% (2/285)	24.1% (7/29)	2.2% (1/46)
Pre-WGD				
K. lactis	n.a.	n.a.	11.5% (3/26)	12.2% (12/98)
D. hansenii	n.a.	n.a.	36% (31/86)	11.4% (21/184)
Y. lipolytica	n.a.	n.a.	1.9% (3/158)	2.9% (21/726)

Notes. Values in brackets indicate the ratio of the number of gene conversions divided by the number of gene comparisons. Data for *S. paradoxus, S. mikatae, S. kudriavzevii, S. bayanus,* and *S. castellii* are not provided because position data was not available for the genes of these genomes. n.a.: not applicable.

Table 4: The number and frequency of gene conversions in ohnologs and paralogs.

Genomes	Ohnologs			Paralogs		
	Number	Frequency (%) with respect to total number of comparisons	Frequency (%) with respect to total number of multigene family members	Number	Frequency (%) with respect to total number of comparisons	Frequency (%) with respect to total number of multigene family members
Post-WGD						
S. cerevisiae	13	2.35	1.17	110	5.71	51.40
S. paradoxus	7	1.60	0.80	44	1.54	9.20
S. mikatae	6	1.45	0.73	26	1.50	4.80
S. kudriavzevii	2	0.88	0.44	20	7.96	29.80
S. bayanus	14	3.03	1.51	50	3.60	12.40
C. glabrata	2	0.67	0.33	8	10.67	14.80
S. castellii	2	0.50	0.25	8	5.06	10.80
Pre-WGD						
K. lactis	n.a.	n.a.	n.a.	15	12.09	23.80
D. hansenii	n.a.	n.a.	n.a.	52	19.25	31.70
Y. lipolytica	n.a.	n.a.	n.a.	24	2.71	8.20

Notes. n.a.: not applicable.

will be 45 ([10(10−1)]/2) whereas it will only be 5 for 10 ohnologs.

Ectopic gene conversions between paralogs are equally frequent in both pre- and post-WGD genomes. Median gene conversion frequencies relative to both total number of comparisons and number of multigene family members are not statistically different between pre-WGD (12.09%, 23.8%) and post-WGD (5.06%, 12.4%) paralogs (Table 4; Wilcoxon two sample test, $P = 0.26$ with respect to the number of gene comparisons and $P = 0.82$ with respect to the number of multigene family members).

There is a significant negative correlation (Spearman rank correlation test) between gene conversion frequency and distance between paralogs located on the same chromosomes in the genomes of S. *cerevisiae* ($r = -0.54$; $P = 0.008$), *C. glabrata* ($r = -0.74$; $P = 0.048$), and *D. hansenii* ($r = -0.45$; $P = 0.008$). Correlations could not be calculated for the other paralog and/or ohnolog data sets either because gene distance information was not available for some species (see above) or because less than four genes were found on the same chromosomes (statistical power analyses require at least 4 data points).

3.4. Gene Conversion Length and Flanking Similarity. The median lengths of the gene conversions between ohnologs are identical in all seven post-WGD genomes (Table 5; multiple comparison ANOVA test, $P = 0.86$, $\alpha = 0.05$). The median lengths of gene conversions between the paralogs of pre-WGD genomes are also equal ($P = 0.34$). However, the median length of the gene conversions between paralogs are significantly longer in *S. cerevisiae* than in *S. paradoxus*, *S. mikatae*, *S. kudriavzevii,* and *S. bayanus* (multiple comparison ANOVA, $P < 0.0001$). In post-WGD genomes, the median length of gene conversion in paralogs and ohnolog (182 and 186.5 bp, resp.) are not significantly different (pairwise Wilcoxon rank tests, Table 5). Finally, the median lengths of gene conversions are significantly different from each other between pre-WGD (150 bp) and post-WGD (182 bp) paralogs (Wilcoxon two sample test, $P = 0.02$,

Table 5: Gene conversion lengths of pre- and post-WGD species.

Genome	Ohnologs (bp)					Paralogs (bp)					Wilcoxon test P-value
	Median	1st quartile	3rd quartile	Min	Max	Median	1st quartile	3rd quartile	Min	Max	
Post-WGD											
S. cerevisiae	272	107	465	60	773	382.5	141	869	8	2642	0.22
S. paradoxus	235	98	354	50	531	106	51.5	232	14	1060	0.17
S. mikatae	165.5	95	431	68	568	167	83	366	14	535	0.64
S. kudriavzevii	270.5	146	395	146	395	136	85	172	25	391	0.19
S. bayanus	149.5	71	315	45	905	126	76	203	21	724	0.50
C. glabrata	83.5	27	140	27	140	130	83.5	386	59	668	0.36
S. castellii	144	118	170	118	170	226	73.5	581.5	44	862	0.69
Pre-WGD											
K. lactis	n.a.	n.a.	n.a.	n.a.	n.a.	99	40	236	32	1127	n.a.
D. hansenii	n.a.	n.a.	n.a.	n.a.	n.a.	183	104.5	310.5	18	1309	n.a.
Y. lipolytica	n.a.	n.a.	n.a.	n.a.	n.a.	83	27.5	196	16	1770	n.a.

Note. Wilcoxon two-sample tests were used to detect differences between the median gene conversion lengths of ohnologs and paralogs. n.a.: not applicable.

Table 6: Maximum flanking similarity of gene conversions in pre and post-WGD species.

Genome	Ohnolog maximum flanking similarity (%)					Paralog maximum flanking similarity (%)					Wilcoxon test P-value
	Median	1st quartile	3rd quartile	Min	Max	Median	1st quartile	3rd quartile	Min	Max	
Post-WGD											
S. cerevisiae	88	84	94	80	97	95.6	91	99	80	100	0.001
S. paradoxus	89	83	94	82	97	90.3	87	97.5	80	100	0.24
S. mikatae	87.5	82	92	81	96	91.7	86.6	95.6	81	100	0.15
S. kudriavzevii	86.8	85.7	88	85.7	88	94	93	97	85	99	0.07
S. bayanus	87.6	85	92	80	98	92.9	86	99	81	100	0.04
C. glabrata	84.5	83	86	83	86	92.6	90	99.5	86	100	0.08
S. castellii	87	86	88	86	88	93	87	97	85	100	0.35
Pre-WGD											
K. lactis	n.a.	n.a.	n.a.	n.a.	n.a.	90	86	97	81	98	n.a.
D. hansenii	n.a.	n.a.	n.a.	n.a.	n.a.	93	86.3	97	80	100	n.a.
Y. lipolytica	n.a.	n.a.	n.a.	n.a.	n.a.	86.5	83.3	93.5	80	100	n.a.

Note. Wilcoxon two sample tests were used to detected differences between the median flanking similarities of ohnologs and paralogs. n.a.: not applicable.

$\alpha = 0.05$). These median lengths are similar to the average length of the *S. cerevisiae* conversions observed in a previous study (173 bp, [4]).

The median sequence similarities of regions flanking gene conversions between ohnologs are equal in all seven post-WGD genomes (Table 6; multiple ANOVA tests, $P = 0.97$, $\alpha = 0.05$). Furthermore, the median sequence similarities of regions flanking gene conversions between paralogs are equal in all seven genomes (multiple comparison ANOVA test, $P = 0.18$, $\alpha = 0.05$).

Although the median flanking similarity of the converted paralogs of post-WGD species is always higher than that of their ohnologs, this difference is only significant in the genome of *S. cerevisiae* and *S. bayanus* (Table 6). However, this lack of statistical significance is likely the result of the relatively low power of these statistical tests because the power of each test was ≤61% (results not shown).

The median sequence similarities of regions flanking gene conversions between the paralogs of pre-WGD genomes are equal (Table 6; multiple ANOVA tests, $P = 0.21$, $\alpha = 0.05$). However, converted genes within pre-WGD paralogs have significantly less flanking similarity (pooled median of 90%) than converted paralogs in post-WGD genomes (pooled median of 94%; Wilcoxon two sample test, $P = 0.0004$, $\alpha = 0.05$, Table 6). We do not know whether this difference has any biological significance.

Analysis of the relationship between the length of gene conversions and flanking similarity indicates a significant positive correlation within the ohnologs of the seven post-WGD genomes (Spearman rank correlation test, $r = 0.44$, $P = 0.005$; Figure 3(a)), the paralogs of the seven post-WGD genomes ($r = 0.36$, $P = 0$; Figure 3(b)) and the paralogs of the three pre-WGD genomes ($r = 0.35$, $P = 0$; Figure 3(c)).

3.5. Ka, Ks, Ka/Ks Ratios and Ontology of Ohnolog and Paralog Converted Genes. In post-WGD genomes, the fact that synonymous substitutions (Ks) are lower for converted paralogs than for converted ohnologs suggests that paralogs have a more recent origin (Table 7). Therefore, the higher Ka/Ks ratio of paralogs clearly indicates that paralogs are under less selection constraints than ohnologs. Furthermore, the similar Ka/Ks ratios of pre- and post-WGD paralogs indicate that the paralogs of pre- and post-WGD evolve under similar selective constraints (Table 7).

The ohnologs and paralogs of *S. cerevisiae* are involved in different processes. Although many of the GO terms shown in Table 8 are not mutually exclusive (e.g., "transposition" and "transposition, RNA-mediated"), analyses of the processes in which these genes are involved show that ohnologs are involved in regulation, essential biosynthetic processes, and metabolic processes whereas paralogs are involved transposition, transport, and nonessential biosynthetic processes.

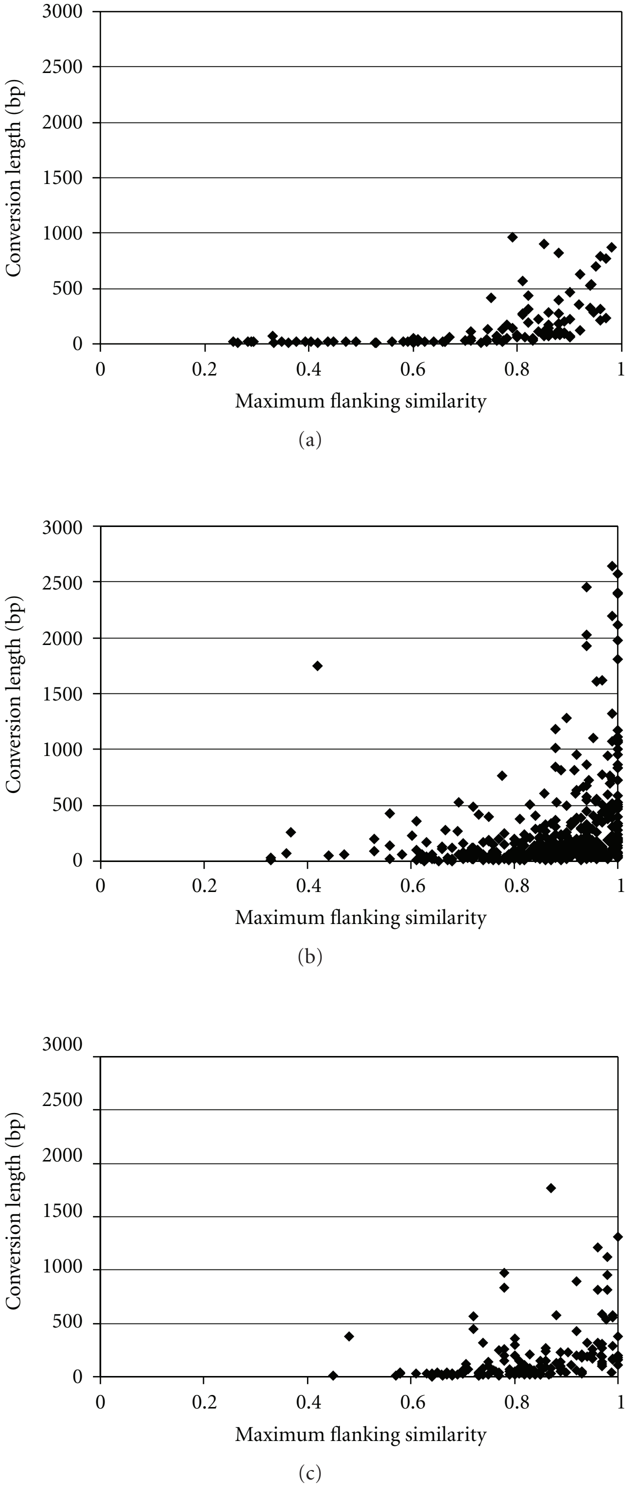
3.6. Location of Converted Regions. When considering pre-WGD paralogs, post-WGD paralogs and post-WGD ohnologs, only the post-WGD paralogs of *S. cerevisiae* show a significant bias of gene conversions towards the 3′-end of genes (Table 9). However, the fact that the power of all nonsignificant tests is smaller than 15% suggests that this bias might also exist in the data sets where it was not detected but that our data are not sufficient to detect it (Table 9).

4. Discussion

Using *S. cerevisiae* ohnologs as queries allowed us to retrieve an average of 372 ohnolog pairs from the other six post-WGD genomes (Table 1). Although these seven species are phylogenetically related (see [13] for a phylogenetic tree of these fungi species), and therefore did not evolve independently, it is very unlikely that species as divergent as *S. cerevisiae* and *C. glabrata* (which diverged soon after the whole genome duplication, some 150 MYA), would have kept 300 pairs of common ohnologs by chance. In fact, assuming that the ancestral pre-WGD genome had 5000 genes and that current post-WGD genomes have 5500 genes [13], one would expect them to have kept only 50 ohnologs in common (0.1 × 0.1 × 5000) by chance alone. As we discuss further below, this suggests that common ohnologs provide a selective advantage and evolve under strong selective constraints.

Since the number and the mean size of paralog multigene families are not significantly different between pre- and post-WGD species, the genome duplication event in the post-WGD genome ancestor did not significantly increase the number or mean size of paralog multigene families in post-WGD species (Table 1, Figure 2). The small number and size of gene families in *C. glabrata* have already been noticed and are likely the result of reductive evolution and gene loss through relatively high genome instability [10, 12, 30].

The chromosomal distribution of ohnologs and paralogs is very different. Whereas, on average, 23.4% of paralogs are found on the same chromosomes, only 4.5% of ohnologs are



(a)

(b)

(c)

Figure 3: Correlation between gene conversion length and maximum flanking sequence similarity. (a) Conversions detected between the ohnologs of the six Saccharomyces species and *C. glabrata*. There are 107 conversions, 46 of which have ≥80% flanking similarity. (b) Conversions detected between the paralogs of the six Saccharomyces species and *C. glabrata*. There are 401 conversions, 311 of which have ≥80% flanking similarity. (c) Conversions detected the paralogs of the three pre-WGD genomes. There are 147 conversions, 91 of which have ≥80% flanking similarity.

Table 7: Nonsynonymous substitutions per nonsynonymous site (Ka), synonymous substitutions per synonymous site (Ks), and Ka/Ks ratios (± standard deviations) for pairs of converted genes in pre- and post-WGD species.

Genome	Ka		Ks		Ka/Ks	
	Ohnologs	Paralogs	Ohnologs	Paralogs	Ohnologs	Paralogs
Post-WGD						
S. cerevisiae	0.04 ± 0.03	0.09 ± 0.08	0.96 ± 0.49	0.37 ± 0.44	0.04 ± 0.02	0.38 ± 0.27
S. paradoxus	0.09 ± 0.11	0.18 ± 0.20	0.91 ± 0.76	0.56 ± 0.40	0.10 ± 0.05	0.46 ± 0.57
S. mikatae	0.09 ± 0.11	0.17 ± 0.19	1.87 ± 1.06	0.56 ± 0.31	0.04 ± 0.04	0.34 ± 0.45
S. kudriavzevii	0.06 ± 0.04	0.08 ± 0.04	0.95 ± 0.46	0.47 ± 0.59	0.06 ± 0.01	0.38 ± 0.34
S. bayanus	0.11 ± 0.09	0.13 ± 0.12	1.91 ± 1.68	0.40 ± 0.46	0.07 ± 0.05	0.40 ± 0.28
C. glabrata	0.25 ± 0.17	0.04 ± 0.04	1.32 ± 0.08	0.36 ± 0.55	0.18 ± 0.12	0.37 ± 0.45
S. castellii	0.18 ± 0.09	0.13 ± 0.07	2.80 ± 1.18	0.29 ± 0.12	0.06 ± 0.01	0.61 ± 0.46
Pre WGD						
K. lactis	n.a.	0.20 ± 0.26	n.a.	0.61 ± 0.40	n.a.	0.49 ± 0.58
D. hansenii	n.a.	0.10 ± 0.07	n.a.	0.50 ± 0.40	n.a.	0.31 ± 0.17
Y. lipolytica	n.a.	0.25 ± 0.19	n.a.	1.12 ± 0.38	n.a.	0.46 ± 0.38

n.a.: not applicable.

Table 8: GO terms associated with biological processes for the ohnologs and paralogs of *S. cerevisiae*.

GO term	Cluster frequency	Background frequency	*P*-value
Ohnologs			
Biological regulation	21.9%	13.8%	5.2×10^{-13}
Regulation of biological process	18.0%	11.3%	3.9×10^{-10}
Regulation of cellular process	16.8%	10.5%	2.6×10^{-09}
External encapsulating structure organization and biogenesis	6.4%	2.8%	6.8×10^{-09}
Cell wall organization and biogenesis	6.4%	2.8%	6.8×10^{-09}
Protein amino acid phosphorylation	4.0%	1.4%	2.1×10^{-08}
Cellular polysaccharide biosynthetic process	2.0%	0.5%	4.1×10^{-08}
Polysaccharide biosynthetic process	2.0%	0.5%	9.9×10^{-08}
Carbohydrate biosynthetic process	2.7%	0.9%	9.3×10^{-07}
Cellular carbohydrate metabolic process	5.2%	2.3%	1.0×10^{-06}
Carbohydrate metabolic process	5.5%	2.5%	1.7×10^{-06}
Paralogs			
Transposition	32.8%	1.3%	9.7×10^{-109}
Transposition, RNA-mediated	32.8%	1.3%	9.7×10^{-109}
Carbohydrate transport	5.2%	0.5%	9.5×10^{-09}
Monosaccharide transport	4.0%	0.3%	4.0×10^{-07}
Hexose transport	4.0%	0.3%	4.0×10^{-07}
Thiamin and derivative metabolic process	3.2%	0.3%	4.0×10^{-05}
Thiamin biosynthetic process	2.8%	0.2%	2.0×10^{-4}
Thiamin and derivative biosynthetic process	2.8%	0.3%	3.1×10^{-4}
Thiamin metabolic process	2.8%	0.3%	3.1×10^{-4}
Telomere maintenance via recombination	2.8%	0.3%	4.8×10^{-4}
Amino acid catabolic process	3.6%	0.5%	1.0×10^{-3}
Cellular response to nitrogen levels	1.6%	0.1%	1.6×10^{-3}

Notes. Frequencies were calculated from 1100 ohnologs, 250 paralogs, and 7159 background genes. Only the twelve most significant results for each type of genes are shown.

Table 9: Correlations between the location of the converted regions and their position in the converted genes in pre- and post-WGD genomes.

Genome	Ohnolog		Paralog	
	R-value	Power	R-value	Power
Post WGD				
S. cerevisiae	−0.07	0.036	0.73*	n.a.
S. paradoxus	0.12	0.049	−0.19	0.072
S. mikatae	0.00	0.025	−0.19	0.076
S. kudriavzevii	−0.17	0.065	−0.09	0.043
S. bayanus	0.24	0.095	0.11	0.047
C. glabrata	0.00	0.025	0.06	0.034
S. castellii	0.17	0.066	−0.09	0.043
Pre WGD				
K. lactis	n.a.	n.a.	−0.32	0.14
D. hansenii	n.a.	n.a.	0.02	0.028
Y. lipolytica	n.a.	n.a.	0.14	0.055

The R-values indicate correlation values. Significant correlations (Spearman rank correlation test $P < 0.05$) are labeled with *. The power of each correlation test is provided except for *S. cerevisiae* paralogs, where the null hypothesis was rejected, and for ohnologs for which a power test could not be performed. n.a.: not applicable.

found on the same chromosomes (Table 2). A likely explanation for this difference is that paralogs often originate from unequal crossing over or replication slippage events whereas ohnologs originate from whole genome duplication events (page 250 of [31], [18], pages 199–202 of [32]). Since gene conversions tend to be more frequent between genes found on the same chromosomes than between genes located on different chromosomes (Table 3), this explains, in part, why gene conversions tend to be more frequent between paralogs than between ohnologs (Table 4). In fact, on average, when comparing gene conversion using total numbers, frequency calculated using the number of multigene family members, or frequency based on the number of gene comparisons, gene conversions are more frequent in the paralogs of pre- and post-WGD genes than in the ohnologs of the post-WGD genomes (Table 4).

The previous work on yeast, Drosophila, and humans has shown that intrachromosomal gene conversions are more frequent than interchromosomal gene conversions [4–6]. A possible explanation for the relatively high frequency of intrachromosomal conversions in *D. hansenii* (36%, Table 3) is that multiple tandem duplication events have been identified within this genome and, therefore, most paralogs are still located on the same chromosomes [10]. In contrast, in *K. lactis* and *Y. lipolytica*, gene conversions between intra- and interchromosomal paralogs are equally frequent (Table 3). The highly redundant *Y. lipolytica* genome has been shown to be undergone a high degree of map dispersion [10]. The low frequency of intrachromosomal conversions observed in this genome might therefore be the result of the dispersion of tandemly duplicated paralogs to other chromosomes. A similar phenomenon might be present in *K. lactis*. It is unlikely that these exceptions are due to mechanistic differences in the repair of double-stranded-breaks between pre- and post-WGD species because the majority of repair genes have been maintained throughout the evolution of the hemiascomycetes [33].

The previous studies have demonstrated a negative correlation between gene conversion frequency and physical distance on the same chromosome [4, 7]. We also observed such a negative correlation in the genomes of *S. cerevisiae*, *C. glabrata,* and *D. hansenii* (see above). However, a lack of data (statistical power) prevented the detection of such a relationship in the paralogs of *K. lactis* and *Y. lipolytica* and the ohnologs of *S. cerevisiae* and *C. glabrata*. This correlation could result from the fact that the DNA repair mechanisms preferentially search for suitable repair templates close to the damaged gene. Since ohnologs are more often found on different chromosomes (Table 2), this would also explain why conversions are less frequent between ohnologs than between paralogs. On the other hand, our recent analyses of the human genome [6] has shown that, in the human genome, the negative correlation between gene conversion frequency and physical distance is simply the result of the fact that most duplicated genes are found next to one another. Thus the negative correlation we observed in some yeast species might also disappear if we normalized our data to take into account the fact that most paralogs are located next to one another on the same chromosome [10].

Sequence similarity requirements for ectopic conversions and the amount of negative selection are very similar between pre- and post-WGD paralogs. Several pieces of information support these conclusions. The fact that the frequency (Table 4), length (Table 5), and flanking sequence similarities (Table 6) of gene conversion of the paralogs within pre- and post-WGD species are similar indicates that mechanistic similarities are present between these genomes. In addition, the fact that the mean Ka/Ks values for the paralog families of pre- and post-WGD species are alike (Table 7) suggests that their genes are under similar selective pressures and have similar gene conversion constraints. This suggests that, despite the different ecological niches of the yeast species, these paralogs evolve in similar ways.

Surprisingly, the sequence of similarity flanking conversions between post-WGD ohnologs is always lower than that flanking post-WGD paralogs (Table 6). This is likely due to the fact that ohnologs are much older than paralogs (i.e., they have larger Ks values; Table 7), which gave time to accumulate more substitutions, and are under more selective constrains (i.e., they have larger Ka/Ks ratios; Table 7). Stronger selective constraints are expected to select against conversions which would homogenize ohnologs because such homogenization would erase the functional differences that each member of a pair of ohnologs has acquired during evolution. As mentioned above, the different function each member of a pair of ohnologs has acquired (neofunctionalization) also likely explains why different yeast genomes have so many common ohnologs (Table 1; [20]). Conversely, one of the effects of repeated gene conversion due to less negative selective pressure on paralogs is that the sequence of similarity between them will increase. Thus, the observation that ectopic gene conversions occur more frequently between paralogs than ohnologs (Table 4) might not only be due to the fact that ohnologs are more often found on different chromosomes (Table 2) but also due to ohnologs being under stronger selective constraints than paralogs (Table 7). These stronger selective constraints are due to the fact that ohnologs are involved in essential processes (regulation, essential biosynthetic processes and metabolic processes) whereas paralogs are involved in nonessential processes (transposition, transport and nonessential biosynthetic processes; Table 8). This is similar to the situation within genes where gene conversions have been shown to be less frequent in more functionally important regions [34, 35].

The previous studies on *S. cerevisiae* have found that gene conversions are biased toward the 3′ end of converted genes. This has been attributed to ectopic gene conversion via cDNA intermediates [4]. Our results confirm that conversions are biased toward the 3′-end of genes within the *S. cerevisiae* paralog dataset [4, Table 9]. The fact that no significant bias was detected within any other species is likely a result of the low statistical power due to the small amount of data available for each of these species (Table 9). This low statistical power for the distribution of gene conversions other than those between *S. cerevisiae* paralogs likely reflects the facts that whereas there were 110 conversions between *S. cerevisiae* paralogs, there were only between 8 and 52 gene conversions between the paralogs of the other nine yeast species (Table 4). They were also only between 2 and 14 gene conversions between the ohnologs of the 7 post-WGD species. These low numbers of gene conversion are therefore not sufficient to ascertain whether their distribution is significantly biased.

The suggestion that the 3′-end bias of the gene conversions between *S. cerevisiae* paralogs is due to ectopic gene conversions with cDNA intermediates is consistent with the low number of introns present in this species as well as their 5′-position bias [4, 36, 37]. The genome of this species contains only 286 introns, and most of these introns are located at the 5′-end of the genes in which they are present [37]. This contrasts with the 139,418 introns found in the human genome and with the absence of intron position bias in human genes [37]. The model proposed by Fink to explain both the paucity and 5′-position bias of *S. cerevisiae* introns posits that incomplete cDNA molecules can recombine with their genomic copies leading to both intron loss and a 5′-position bias of the remaining introns [36, 37]. This model was later supported by the experimental demonstration that cDNA molecule can recombine with their genomic copy [9]. Since the genomes of *C. glabrata*, *D. hansenii*, *K. lactis,* and *Y. lipolytica* all have few introns and that their introns have a 5′-position bias [38], one would also expect to observe a 3′-end bias for their gene conversions if they often occur with cDNA copies. As discussed above, the fact that we did not observe such a bias in these four species could be due to the low statistical power of our tests. Alternatively, it could reflect recombination differences between *S. cerevisiae* and these four species.

In summary, our results show that the number and mean size of multigene families composed of paralogous sequences are not significantly different between pre- and post-WDG species (Table 1, Figure 2), that paralogs are more often found on the same chromosomes than ohnologs (Table 2), that gene conversions tend to be more frequent between genes found on the same chromosomes than between genes located on different chromosomes (Table 3), that gene conversions tend to be more frequent between paralogs than between ohnologs (Table 4), that the frequency (Table 4), length (Table 5), and flanking sequence similarities (Table 6) of the gene conversions between the paralogs of pre- and post-WGD species are similar, that there is a positive correlation between the length of gene conversions and flanking similarity in all converted genes (Figure 3), that ohnologs are under stronger selective constraints than paralogs (Table 7), that these stronger selective constraints are due to the fact that ohnologs are involved in essential processes whereas paralogs are involved in nonessential processes (Table 8), and that conversions are biased toward the 3′-end of the *S. cerevisiae* paralogs (Table 9). In the future, since it has recently been shown that the expression levels of duplicated genes influence their rate of sequence divergence [39], it would be interesting to test whether the increased ectopic gene conversion frequency we observed in *C. glabrata*, *D. hansenii*, and *K. lactis* (Table 3) is due to conversions between highly expressed genes.

Acknowledgments

The authors thank the two anonymous referees for their useful and constructive comments on a previous version of this paper. This work was supported by a Discovery Grant from the Natural Science and Engineering Research Council of Canada to G. Drouin.

References

[1] Y. Aylon and M. Kupiec, “DSB repair: the yeast paradigm,” *DNA Repair*, vol. 3, no. 8-9, pp. 797–815, 2004.

[2] V. M. Watt, C. J. Ingles, M. S. Urdea, and W. J. Rutter, “Homology requirements for recombination in *Escherichia coli*,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 82, no. 14, pp. 4768–4722, 1985.

[3] P. Shen and H. V. Huang, “Homologous recombination in *Escherichia coli*: dependence on substrate length and homology,” *Genetics*, vol. 112, no. 3, pp. 441–457, 1986.

[4] G. Drouin, “Characterization of the gene conversions between the multigene family members of the yeast genome,” *Journal of Molecular Evolution*, vol. 55, no. 1, pp. 14–23, 2002.

[5] W. R. Engels, C. R. Preston, and D. M. Johnson-Schlitz, “Long-range *cis* preference in DNA homology search over the length of a Drosophila chromosome,” *Science*, vol. 263, no. 5153, pp. 1623–1625, 1994.

[6] D. Benovoy and G. Drouin, “Ectopic gene conversions in the human genome,” *Genomics*, vol. 93, no. 1, pp. 27–32, 2009.

[7] A. S. H. Goldman and M. Lichten, “The efficiency of meiotic recombination between dispersed sequences in *Saccharomyces cerevisiae* depends upon their chromosomal location,” *Genetics*, vol. 144, no. 1, pp. 43–55, 1996.

[8] G. Achaz, E. Coissac, A. Viari, and P. Netter, “Analysis of intrachromosomal duplications in yeast *Saccharomyces*

cerevisiae: a possible model for their origin," *Molecular Biology and Evolution*, vol. 17, no. 8, pp. 1268–1275, 2000.

[9] L. K. Derr and J. N. Strathern, "A role for reverse transcripts in gene conversion," *Nature*, vol. 361, no. 6408, pp. 170–173, 1993.

[10] B. Dujon, D. Sherman, G. Fischer et al., "Genome evolution in yeasts," *Nature*, vol. 430, no. 6995, pp. 35–44, 2004.

[11] K. H. Wolfe and D. C. Shields, "Molecular evidence for an ancient duplication of the entire yeast genome," *Nature*, vol. 387, no. 6634, pp. 708–713, 1997.

[12] M. Kellis, B. W. Birren, and E. S. Lander, "Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*," *Nature*, vol. 428, no. 6983, pp. 617–624, 2004.

[13] K. Wolfe, "Evolutionary genomics: yeasts accelerate beyond BLAST," *Current Biology*, vol. 14, no. 10, pp. R392–R394, 2004.

[14] D. R. Scannell, K. P. Byrne, J. L. Gordon, S. Wong, and K. H. Wolfe, "Multiple rounds of speciation associated with reciprocal gene loss in polyploid yeasts," *Nature*, vol. 440, no. 7082, pp. 341–345, 2006.

[15] A. Goffeau, G. Barrell, H. Bussey et al., "Life with 6000 genes," *Science*, vol. 274, no. 5287, pp. 546–567, 1996.

[16] P. Cliften, P. Sudarsanam, A. Desikan et al., "Finding functional features in Saccharomyces genomes by phylogenetic footprinting," *Science*, vol. 301, no. 5629, pp. 71–76, 2003.

[17] M. Kellis, N. Patterson, M. Endrizzi, B. Birren, and E. S. Lander, "Sequencing and comparison of yeast species to identify genes and regulatory elements," *Nature*, vol. 423, no. 6937, pp. 241–254, 2003.

[18] K. H. Wolfe, "Yesterday's polyploids and the mystery of diploidization," *Nature Reviews Genetics*, vol. 2, no. 5, pp. 333–341, 2001.

[19] A. van Hoof, "Conserved functions of yeast genes support the duplication, degeneration and complementation model for gene duplication," *Genetics*, vol. 171, no. 4, pp. 1455–1461, 2005.

[20] S. Wong and K. H. Wolfe, "Duplication of genes and genomes in yeasts," in *Comparative Genomics*, P. Sunnerhagen and J. Piskur, Eds., vol. 15, pp. 78–99, Springer, Heidelberg, Germany, 2005.

[21] K. P. Byrne and K. H. Wolfe, "The Yeast Gene Order Browser: combining curated homology and syntenic context reveals gene fate in polyploid species," *Genome Research*, vol. 15, no. 10, pp. 1456–1461, 2005.

[22] J. D. Thompson, D. G. Higgins, and T. J. Gibson, "CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice," *Nucleic Acids Research*, vol. 22, no. 22, pp. 4673–4680, 1994.

[23] S. Sawyer, GENECONV molecular biology computer program, 1999, http://www.math.wustl.edu/~sawyer/geneconv.

[24] D. Posada and K. A. Crandall, "Evaluation of methods for detecting recombination from DNA sequences: computer simulations," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 24, pp. 13757–13762, 2001.

[25] E. Erdfelder, F. Faul, and A. Buchner, "GPOWER: a general power analysis program," *Behavior Research Methods, Instruments, & Computers*, vol. 28, no. 1, pp. 1–11, 1996.

[26] Z. Yang, "PAML: a program package for phylogenetic analysis by maximum likelihood," *CABIOS*, vol. 13, no. 5, pp. 555–556, 1997.

[27] Z. Yang and R. Nielsen, "Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models," *Molecular Biology and Evolution*, vol. 17, no. 1, pp. 32–43, 2000.

[28] E. L. Hong, R. Balakrishnan, Q. Dong et al., "Gene Ontology annotations at SGD: new data sources and annotation methods," *Nucleic Acids Research*, vol. 36, no. 1, pp. D577–D581, 2008.

[29] W. P. Rice, "Analysing tables of statistical tests," *Evolution*, vol. 43, no. 1, pp. 223–225, 1989.

[30] G. Fischer, E. P. Rocha, F. Brunet, M. Vergassola, and B. Dujon, "Highly variable rates of genome rearrangements between hemiascomycetous yeast lineages," *PLoS Genetics*, vol. 2, no. 3, article e32, 2006.

[31] D. Graur and W.-H. Li, *Fundamentals of Molecular Evolution*, Sinauer Associates, Sunderland, Mass, USA, 2nd edition, 2000.

[32] M. Lynch, *The Origins of Genome Architecture*, Sinauer Associates, Sunderland, Mass, USA, 2007.

[33] G.-F. Richard, A. Kerrest, I. Lafontaine, and B. Dujon, "Comparative genomics of hemiascomycete yeasts: genes involved in DNA replication, repair, and recombination," *Molecular Biology and Evolution*, vol. 22, no. 4, pp. 1011–1023, 2005.

[34] Z. Zhao, D. Hewett-Emmett, and W.-H. Li, "Frequent gene conversion between human red and green opsin genes," *Journal of Molecular Evolution*, vol. 46, no. 4, pp. 494–496, 1998.

[35] J. P. Noonan, J. Grimwood, J. Schmutz, M. Dickson, and R. M. Myers, "Gene conversion and the evolution of protocadherin gene cluster diversity," *Genome Research*, vol. 14, no. 3, pp. 354–366, 2004.

[36] G. R. Fink, "Pseudogenes in yeast?" *Cell*, vol. 49, no. 1, pp. 5–6, 1987.

[37] T. Mourier and D. C. Jeffares, "Eukaryotic intron loss," *Science*, vol. 300, no. 5624, p. 1393, 2003.

[38] D.-K. Niu, W.-R. Hou, and S.-W. Li, "mRNA-mediated intron losses: evidence from extraordinarily large exons," *Molecular Biology and Evolution*, vol. 22, no. 6, pp. 1475–1481, 2005.

[39] S. Pyne, S. Skiena, and B. Futcher, "Copy correction and concerted evolution in the conservation of yeast genes," *Genetics*, vol. 170, no. 4, pp. 1501–1513, 2005.

SAGE-Hindawi Access to Research
International Journal of Evolutionary Biology
Volume 2011, Article ID 921312, 11 pages
doi:10.4061/2011/921312

Research Article

Sequence Analysis of SSR-Flanking Regions Identifies Genome Affinities between Pasture Grass Fungal Endophyte Taxa

Eline van Zijll de Jong,[1, 2, 3] Kathryn M. Guthridge,[1, 2, 4] German C. Spangenberg,[1, 2, 4, 5] and John W. Forster[1, 2, 4, 5]

[1] *Department of Primary Industries, Biosciences Research Division, Victorian AgriBiosciences Centre, 1 Park Drive, La Trobe University Research and Development Park, Bundoora, VIC 3083, Australia*
[2] *Molecular Plant Breeding Cooperative Research Centre, La Trobe University Research and Development Park, Bundoora, VIC 3083, Australia*
[3] *Bioprotection Research Centre, P.O. Box 84, Lincoln University, Lincoln 7647, Canterbury, New Zealand*
[4] *Dairy Futures Cooperative Research Centre, La Trobe University Research and Development Park, Bundoora, VIC 3083, Australia*
[5] *La Trobe University, Bundoora, VIC 3086, Australia*

Correspondence should be addressed to John W. Forster, john.forster@dpi.vic.gov.au

Received 14 October 2010; Accepted 10 December 2010

Academic Editor: Hiromi Nishida

Fungal species of the *Neotyphodium* and *Epichloë* genera are endophytes of pasture grasses showing complex differences of life-cycle and genetic architecture. Simple sequence repeat (SSR) markers have been developed from endophyte-derived expressed sequence tag (EST) collections. Although SSR array size polymorphisms are appropriate for phenetic analysis to distinguish between taxa, the capacity to resolve phylogenetic relationships is limited by both homoplasy and heteroploidy effects. In contrast, nonrepetitive sequence regions that flank SSRs have been effectively implemented in this study to demonstrate a common evolutionary origin of grass fungal endophytes. Consistent patterns of relationships between specific taxa were apparent across multiple target loci, confirming previous studies of genome evolution based on variation of individual genes. Evidence was obtained for the definition of endophyte taxa not only through genomic affinities but also by relative gene content. Results were compatible with the current view that some asexual *Neotyphodium* species arose following interspecific hybridisation between sexual *Epichloë* ancestors. Phylogenetic analysis of SSR-flanking regions, in combination with the results of previous studies with other EST-derived SSR markers, further permitted characterisation of *Neotyphodium* isolates that could not be assigned to known taxa on the basis of morphological characteristics.

1. Introduction

Fungal endophytes of the genus *Neotyphodium* and *Epichloë* are widespread in temperate grasses of the Poaceae subfamily Pooideae [1, 2]. In agronomically important pasture grasses such as perennial ryegrass (*Lolium perenne* L.) and tall fescue (*Festuca arundinacea* Schreb. [Darbysh.] syn. *L. arundinaceum*), the respective symbionts (*N. lolii* [Latch, Christensen, and Samuels] Glenn, Bacon, and Hanlin and *N. coenophialum* [Morgan-Jones and Gams] Glenn, Bacon, and Hanlin) confer both beneficial and detrimental agronomic traits [3–7]. Molecular genetic marker-based studies have contributed to knowledge of endophyte genetics, taxonomy, and phylogeny. *Neotyphodium* species were originally placed in the form (asexual) genus *Acremonium* [8], but were reclassified into the new form genus *Neotyphodium* when sequence analysis of ribosomal RNA-encoding (rDNA) genes indicated a monophyletic group with the sexual *Epichloë* species [1, 9]. Phylogenetic analysis of genes for conserved proteins such as the β-tubulin gene (*tub2*), translation elongation factor 1-α (*tefA*), and actin (*actG*) also provided evidence for close relationships between *Neotyphodium* and *Epichloë* species [10–17].

Although taxa such as *N. lolii* are haploid in nature, other *Neotyphodium* species were shown to contain multiple gene copies and conform to heteroploid genomic constitutions [17]. The single or multiple gene copies of asexual *Neotyphodium* species appear to correspond to those of specific haploid *Epichloë* species. This observation has been interpreted to support a hybrid origin for heteroploid taxa: for instance, *N. coenophialum* has been proposed to have arisen through hybridisation and subsequent nuclear fusion events involving the extant taxa *E. typhina*, *E. baconii,* and *E. festucae.* The relative genome sizes of haploid and heteroploid endophytes (c. 30 Mb for *N. lolii*; c. 60 Mb for *N. coenophialum*) lend some support to this hypothesis [18], subject to the possibility of selective gene loss subsequent to hybridisation events. Phylogenetic relationships between endophyte taxa are hence complex and reticulated. Sequence analysis of individual gene loci may be used to infer such relationships based on affinities between shared genomes. However, performance differences between individual genes have been observed. The resolution capacity provided by rDNA and *actA* genes was low in comparison to other genes [12–15], possibly due to homoplasy effects [13]. Heteroploid-like *Neotyphodium* species also display aneuploidy for some loci, such as the rDNA gene, limiting resolution of complete phylogenies [19]. A broader survey of gene classes is hence desirable to further clarify affinities between endophyte taxa.

Simple sequence repeats (SSRs) or microsatellites [20] have been widely used for analysis of genetic variation within and between closely related species [21]. A high rate of mutation [22] renders SSR array length polymorphism particularly useful for intraspecific genetic studies. However, sequence analysis has revealed complex mechanisms controlling allele size variation, limiting the efficiency of interspecific phylogenetic analysis. Repeat number variation is thought to arise from polymerase slippage during replication [23], but constraints on threshold size for allele expansion [24] and on allele size range [25] are evident. In addition, interruptions of the repeat structure tend to stabilise SSR loci [26]. Constraints on allele size may consequently lead to inaccurate assessment of phylogenetic divergence between taxa. Size homoplasy of distinct alleles arising from insertions, deletions, and base substitutions in the SSR flanking regions are also common [27–30]. Changes in flanking regions appear to occur independently of changes in the SSR repeat array [28, 30]. Due to these factors, allelic variation of SSR loci, as assessed by amplicon size variation, is appropriate only for phenetic analysis and not suitable for phylogenetic reconstruction.

In contrast, several studies have performed phylogenetic interpretation through analysis of SSR-flanking sequence regions. The resolving power of evolutionary studies using individual structural genes may be constrained by limited divergence [31], and studies of a small number of gene loci may not be representative of whole-genome variation [32]. However, the abundant genomic distribution of SSRs [33–35] permits phylogenetic assessment across the transcriptional units of multiple gene classes. SSR-flanking regions have been used for phylogenetic analysis of multiple organisms [31, 32, 36, 37], resolving relationships to otherwise inaccessible levels [31, 32, 37].

Consistent with these previous studies, gene-associated SSR loci have previously been shown to discriminate endophyte taxa based on size polymorphism [38], but did not permit phylogenetic analysis. The present study describes the comparison of sequences that flank the SSR array in 5 independently selected gene loci, across 23 distinct fungal endophyte isolates. The derived data have determined the extent of molecular variation underlying SSR size polymorphism, confirmed current models for genome affinities, inferred phylogenetic relationships and models of genome evolution (including a role for selective gene loss), and elucidated the genomic origin of several previously unclassified *Neotyphodium* taxa.

2. Materials and Methods

2.1. Endophyte Isolates. Phylogenetic analysis was performed on 20 endophyte isolates representing three *Neotyphodium* and five *Epichloë* taxa, as well as three *Neotyphodium* isolates which could not be assigned to known taxa based on their morphological characters (A. Leuchtmann, pers. comm.), and a tall fescue endophyte taxon (FaTG-2) which has yet to be allocated a Linnean name (Table 1). Endophyte isolates were cultured and DNA was extracted as described previously [38].

2.2. DNA Sequence Analysis of EST-SSR Amplicons. Genomic amplicons obtained with primer pairs designed to five EST-SSR loci (NCESTA1AB04, NCESTA1FH03, NCESTA1AG07, NLESTA1GF09, and NLESTA1NF04) were analysed. Amplicons were obtained as described previously [38]. Amplicons from haploid taxa were analysed by direct sequencing, with the exception of locus NLESTA1NF04 for which sequencing was performed on purified plasmids containing the cloned amplicon [38]. Sequencing reactions were performed in 10 μl reaction volumes containing 4 μl Sequencing Reagent Premix from the DYEnamic ET Terminator Cycle Sequencing kit (Amersham Biosciences, Little Chalfont, UK), 0.5 μM of forward or reverse primer for the locus of interest and 5 μl of amplicon in a thermocycler (GeneAmp; PE Applied Biosystems, Forster City, California, USA.) programmed for 20 seconds at 92°C followed by 30 cycles of 20 s at 95°C, 15 s at 50°C, 2 minutes at 60°C, then 10 min at 60°C. Sequencing products were purified using Autoseq 96 plates (Amersham Biosciences), dried at 80°C for 30 min and resuspended in 5 μl of sterile Milli-Q water before analysis on a ABI Prism 3700 automated sequencer (PE Applied Biosystems). For multiple products amplified in a single reaction from nonhaploid taxa, cloning was used to separate the different amplicons. Following amplification, the products were purified using a Microspin S-300 HR Column (Amersham Biosciences). The purified products were cloned into pGEM-T Easy Vector (Promega, Madison, Wisconsin,

Table 1: Endophyte isolates used for phylogenetic analysis.

Species or taxon	Isolate	Host species	Origin	Source
N. coenophialum	9309	*Festuca arundinacea*	France	ETH Zürich[1]
N. coenophialum	9920/1	*F. arundinacea*	U.S.A.	ETH Zürich
N. coenophialum	9920/2	*F. arundinacea*	U.S.A.	ETH Zürich
N. coenophialum	9920/3	*F. arundinacea*	U.S.A.	ETH Zürich
FaTG-2	8907	*F. arundinacea*	U.S.A.	ETH Zürich
N. lolii	9601	*Lolium perenne*	Belgium	ETH Zürich
N. lolii	Ellett H5837	*L. perenne*	New Zealand	DPI-Hamilton[2]
N. lolii	North African 6	*L. perenne*	Morocco	DPI-Hamilton
N. lolii	Victorian 2	*L. perenne*	Australia	DPI-Hamilton
N. uncinatum	9414	*F. pratensis*	Germany	ETH Zürich
Neotyphodium sp.	9303/2	*Elymus europaeus*	Switzerland	ETH Zürich
Neotyphodium sp.	9727	*F. arizonica*	U.S.A.	ETH Zürich
Neotyphodium sp.	9728	*L. perenne*	New Zealand	ETH Zürich
E. baconii	9707	*Agrostis tenuis*	Switzerland	ETH Zürich
E. bromicola	9630	*Bromus erectus*	Switzerland	ETH Zürich
E. clarkii	9401	*Holcus lanatus*	Switzerland	ETH Zürich
E. festucae	9412	*F. gigantea*	Switzerland	ETH Zürich
E. festucae	9436	*F. pratensis*	Switzerland	ETH Zürich
E. festucae	9713	*F. rubra*	Switzerland	ETH Zürich
E. festucae	9718	*F. gigantea*	Switzerland	ETH Zürich
E. festucae	9722	*F. rubra*	England	ETH Zürich
E. sylvatica	9301	*Brachypodium sylvaticum*	Switzerland	ETH Zürich
E. typhina	9635	*Dactylis glomerata*	Switzerland	ETH Zürich

[1] ETH Zürich: Geobotanisches Institut, ETH, Zürich, Switzerland.
[2] DPI-Hamilton: Department of Primary Industries, Primary Industries Research Victoria, Hamilton, Victoria, Australia.

U.S.A.) and transformed into competent cells. Inserts were amplified from transformed colonies and sequenced.

Consensus sequences were derived through analysis of several independently isolated clones or direct sequencing of both strands. Sequences were compared using Sequencher (version 4.0) (Gene Codes Corporation, Ann Arbor, Michigan, U.S.A.). BLASTX (version 2.2.1 and 2.2.6) [39] was used to search for similarities between the EST sequence of SSR loci and protein sequences in the protein databases available from the National Centre for Biological Information (NR, PDB and SwissProt; http://www.ncbi.nlm.nih.gov/BLAST/).

2.3. Phylogenetic Analysis of EST-SSR Amplicons. The DNA sequences of unique amplicons were prepared for phylogenetic analysis by compilation in FastA format into a single file for sequence alignment in ClustalX (version 1.8) [40]. Manual realignment of sequences removed primer termini and polymorphic SSR arrays and converted insertion-deletion (indel) regions into single multistate characters. Sequence alignments were analysed by clustering or tree searching methods available in PHYLIP (version 3.6a3) (J. Felsenstein, University of Washington, Seattle, Washington, U.S.A., available from http://evolution.gs.washington.edu/phylip.html). For parsimony analysis, sequences were analysed with indels either removed, or coded as single multistate characters. Where multiple trees were resolved, the Kishino-Hasegawa-Templeton (KHT) test [41, 42] or Shimodaira-Hasegawa (SH) test [43] was used to test for significant differences. The robustness of the trees was measured by the Bootstrap method [44] with 1000 replicates. A bootstrap value of 70% or greater was considered to be well supported. For Maximum Likelihood (ML) and distance-based analysis, sequences were analysed with indels removed. Distance matrices were obtained using the F84 model [42, 45] and clustered using the Fitch-Margoliash (FM) method [46] or Neighbor-Joining (NJ) method [47]. The transition/transversion ratio was estimated using the Tree-Puzzle program (version 5.0) [48] or by the ML method. To estimate the transition/transversion ratio by the ML method, different possible values for the transition/transversion ratio were evaluated in multiple runs to find the value with the maximum likelihood estimate. The same approach was used to estimate the among-site rate heterogeneity by the ML method. To estimate the among-site rate heterogeneity by the Minimum Evolution (ME) method, distance matrices generated for each site were analysed and the total branch length was taken as the estimated value of the rate of change for each of the sites.

3. Results

3.1. Characterisation of Endophyte EST-SSR Amplicons. Genomic amplicons from five EST-SSR loci, ranging in length from 181–385 bp, were characterised from 12

Table 2: Characteristics of EST-derived SSR loci used for phylogenetic analysis of endophyte isolates.

	EST-SSR locus				
	NCESTA1AB04	NCESTA1FH03	NCESTA1GA07	NLESTA1GF09	NLESTA1NF04
Product size (bp)	241–257	217–252	181–206	364–385	242–265
Number of indels	5	7	3	6	2
Number of unique SSR array variants	2	13	11	1	19
Number of unique gene variants	14	13	13	11	15
Composition of products	coding + 5′-UTR	unknown	intron + unknown	coding + intron	unknown
Length of unit for phylogenetic analysis[1] (bp)	195	163	116	321	166
CDS or exon	151		46	238	
UTR or intron	44		70	83	
Number of informative characters[1]	40	37	23	44	51
CDS or exon	30		10	30	
UTR or intron	10		13	14	
Percent informative characters[1]	21	21	20	14	31
CDS or exon	20		22	13	
UTR or intron	23		19	17	
Number of trees resolved	3	1	4	2	2

[1] Each indel was coded as a single multistate character.

Neotyphodium and 10 *Epichloë* isolates, as well as FaTG-2 (Table 2). The sequences of two loci (NCESTA1AB04 and NLESTA1GF09) shared amino acid sequence similarity with hypothetical or predicted proteins of *Neurospora crassa* and *Magnoporthe grisea* and were mainly composed of coding sequence (CDS) as well as 5′-untranslated region (5′-UTR) and intron sequences, respectively. The sequences of the remaining three loci did not show similarity with any proteins in public databases. Amplification products from locus NCESTA1GA07 were predominantly composed of intron sequence, based on comparison of EST and genomic DNA sequences.

Size polymorphisms between taxa for the selected loci resulted from variation at a number of indel sites (Table 2). Differences also occurred in the repeat unit number of the SSR array for three loci (NCESTA1FH03, NCESTA1GA07 and NLESTA1NF04; data not shown), accounting for the majority of the observed size polymorphisms. A number of sequence haplotypes (defined here as amplicons with multiple sequence variant content, but clearly related to a single common reference) for each locus were identified across the sample set, with identity observed between those of several *Neotyphodium* and *Epichloë* isolates. Single haplotypes for individual genes were observed for *N. lolii*, unclassified *Neotyphodium* isolate 9727, and the different *Epichloë* species, while *N. coenophialum*, FaTG-2, *N. uncinatum,* and unclassified *Neotyphodium* isolates 9303/2 and 9728 generated multiple haplotypes. The number of haplotypes present in these species varied between loci, but with a maximum of three for *N. coenophialum* and *Neotyphodium* isolates 9303/2 and 9728, and two for FaTG-2 and *N. uncinatum*. Variation was observed in cloning efficiency of different PCR products for those species possessing multiple haplotypes. Aberrant haplotypes, which were likely to be chimeras generated by PCR-mediated recombination, were also obtained.

Pairwise comparisons identified between 23–51 informative characters for the different loci (Table 2). The proportion of informative characters ranged from 14% (locus NLESTA1GF09) to 31% (locus NLESTA1NF04), but was c. 20% for the other loci. A similar proportion of informative characters occurred in both the coding and noncoding sequences of the eligible loci.

3.2. Phylogenetic Analysis of Endophyte EST-SSR Amplicons.

3.2. Phylogenetic Analysis of Endophyte EST-SSR Amplicons. Loci were analysed through sequence alignment (Supplementary Material: Appendices 1–5 available at doi:10.4067/2011/921312) individually, rather than as a combined dataset, due to variation of both inferred ploidy level and number of observed haplotypes between different SSR loci from heteroploid isolates. Between one and four trees were resolved for the different loci using the Parsimony method (Figures 1–5). A single tree was resolved for locus NCESTA1FH03 (Figure 2). The multiple trees obtained for loci NCESTA1AB04 (Figure 1), NLESTA1GF09 (Figure 4) and NLESTA1NF04 (Figure 5) only differed in the placement of one or two species. More variation was evident in the branching of the multiple trees identified for the locus NCESTA1GA07 (Figure 3). The trees, however, were not found to be significantly different in the KHT or SH tests. The majority of branches in the trees were supported by bootstrap analysis. Similar trees were resolved for the different loci using the ML, FM, and NJ methods (data not shown). In tests performed using the ML and ME methods, no significant differences were detected in the rate of change between coding and non-coding sequences or between exon and intron sequences for eligible loci (data not shown).

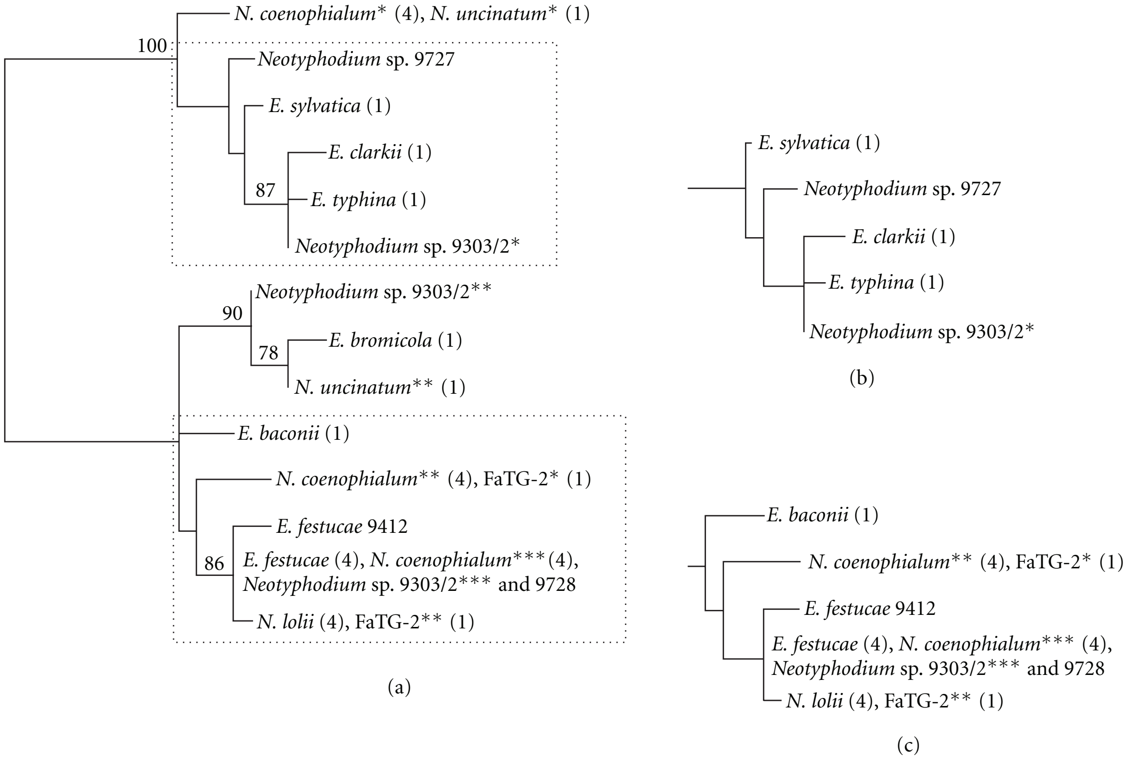

N. coenophialum* (4), N. uncinatum* (1)
100
Neotyphodium sp. 9727
E. sylvatica (1)
E. clarkii (1)
87
E. typhina (1)
Neotyphodium sp. 9303/2*
90
Neotyphodium sp. 9303/2**
78
E. bromicola (1)
N. uncinatum** (1)
E. baconii (1)
N. coenophialum** (4), FaTG-2* (1)
E. festucae 9412
86
E. festucae (4), N. coenophialum*** (4), Neotyphodium sp. 9303/2*** and 9728
N. lolii (4), FaTG-2** (1)
(a)
E. sylvatica (1)
Neotyphodium sp. 9727
E. clarkii (1)
E. typhina (1)
Neotyphodium sp. 9303/2*
(b)
E. baconii (1)
N. coenophialum** (4), FaTG-2* (1)
E. festucae 9412
E. festucae (4), N. coenophialum*** (4), Neotyphodium sp. 9303/2*** and 9728
N. lolii (4), FaTG-2** (1)
(c)

Figure 1: Parsimony analysis of sequence haplotypes derived from reference *Neotyphodium* and *Epichloë* isolates for the EST-SSR locus NCESTA1AB04. (a) One of three most parsimonious trees obtained. The left edge is the inferred midpoint root. Branches with bootstrap values of greater than 70% from 1000 bootstrap replications are marked. The number of isolates with an identical haplotype is indicated in the brackets following the species name. In the instances when multiple haplotypes were identified, the relevant variant number is indicated by the number of asterisks following the species name or isolate number. The boxed area indicated by the dotted line indicates the regions representing subtrees. ((b) and (c)) Subtrees showing the alternative branching of *E. sylvatica* and *E. baconii* in the other trees.

Phylogenetic analysis of the different loci revealed similar genomic relationships among endophyte species, as summarised in a network format (Figure 6). Close relationships between haplotypes from different taxa were deduced to indicate partial or complete commonality of genome content. Some locus-dependent differences in tree topology (Figures 1–5) were observed, but specific taxa consistently grouped together. In most instances, the *Epichloë* species were separated into two groups, separation being supported by bootstrap analysis. The first contained *E. festucae*, *E. baconii,* and *E. bromicola*, and the second contained *E. typhina*, *E. clarkii,* and *E. sylvatica*, *Neotyphodium* species being included within both groups. Group 1 *Epichloë* species were further divided into distinct branches according to their taxonomic classification, while Group 2 endophytes showed higher levels of genetic similarity. Close genetic relationships were evident between several *Neotyphodium* and *Epichloë* species. These relationships were observed in most of the trees and were also supported by bootstrap analysis. The single haplotypes derived from both *N. lolii* and *E. festucae* were grouped together in all trees and were identical in structure for two of the loci (NCESTA1FH03 and NLESTA1GF09). Multiple haplotypes from *N. coenophialum*, FaTG-2, and *N. uncinatum* were consistently associated with counterparts from specific *Neotyphodium* and *Epichloë* species. *N. coenophialum* and *N. uncinatum* shared identical or very closely related haplotypes for all five loci. These variants also grouped with those from Group 2 *Epichloë* species. *N. coenophialum* also shared common haplotypes with FaTG-2 and *E. baconii.* The remaining haplotypes common to *N. coenophialum* and FaTG-2 grouped with the corresponding single haplotypes from *E. festucae* and *N. lolii.* For a subset of the target loci, *N. uncinatum*-derived sequences grouped with the corresponding haplotypes from *E. bromicola*.

Unclassified *Neotyphodium* isolates also displayed close genetic relationships with known taxa. *Neotyphodium* isolate 9727 produced single haplotypes from each locus that were either identical or very similar to the haplotypes common between *N. coenophialum* and *N. uncinatum*, and grouped most closely with those derived from *E. sylvatica*. *Neotyphodium* isolates 9303/2 and 9728 were closely related, all locus-specific haplotypes showing a high degree of sequence similarity. One of three subclasses of derived haplotypes grouped to form a distinct well-supported group with those from *E. festucae*, *N. lolii*, *N. coenophialum,* and FaTG-2. Isolates 9303/2 and 9728 exhibited a second haplotype subclass that grouped with counterparts from *E. bromicola* and *N. uncinatum,* and the same class was present in isolate over two loci. The remaining haplotype sub-class (observed in isolate 9303/2 for four loci, and in isolate 9728 for one locus)

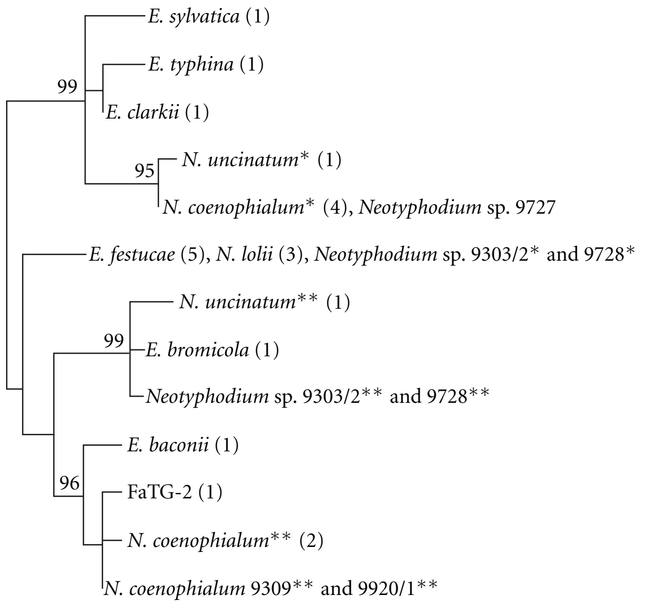

E. sylvatica (1)
E. typhina (1)
99
E. clarkii (1)
95
N. uncinatum* (1)
N. coenophialum* (4), Neotyphodium sp. 9727
E. festucae (5), N. lolii (3), Neotyphodium sp. 9303/2* and 9728*
N. uncinatum** (1)
99
E. bromicola (1)
Neotyphodium sp. 9303/2** and 9728**
E. baconii (1)
96
FaTG-2 (1)
N. coenophialum** (2)
N. coenophialum 9309** and 9920/1**

Figure 2: Parsimony analysis of sequence haplotypes derived from reference *Neotyphodium* and *Epichloë* isolates for the EST-SSR locus NCESTA1FH03. Diagram properties are as described in the legend to Figure 1. Note that the nucleotide sequence of the amplification product from *N. lolii* isolate North African 6 detected at this locus by autoradiography [38] was not obtained.

grouped with the equivalent haplotypes from *E. typhina* and *E. clarkii*.

4. Discussion

4.1. Application of EST-SSR Loci to Endophyte Phylogenetic Analysis. The application of SSR markers for phylogenetic analysis is limited by two main factors: complex molecular evolution of SSR loci and the occurrence of size homoplasy between distinct SSR alleles. Sequence analysis of selected SSR loci in the current study has demonstrated that these factors influence the generic inability of SSR markers to resolve phylogenetic relationships among endophyte species [38]. Changes in the SSR array repeat number appeared to be independent of flanking region changes: some of the locus NLESTA1NF04-derived haplotypes from multiple *E. festucae* isolate differed for SSR array number, but exhibited identity for flanking sequence, while others showed the converse relationship. SSR allele size homoplasy occurred between different endophyte taxa of distinct origins as a result of insertions, deletions, and base substitutions in both the SSR motif and flanking sequences, as observed for the locus NCESTA1FH03-specific *E. festucae* and *E. baconii*-related *N. coenophialum* haplotypes. Endophyte SSR locus arrays were highly variable, and differences in repeat unit number generally accounted for allele size variation between closely related endophyte species, while indel and base substitution incidence increased when comparisons were made between more distantly related taxa.

The results of this and other studies [30, 31] suggest that size variation may provide a relatively accurate measure of genetic variation between closely related species. Although homoplasy was not taken into account, SSRs have previously proven useful for genetic discrimination within and between endophyte species [38]. Presumably, the inherently variable nature of SSRs and the large number of loci analysed reduced the potential biasing effects of individual loci. The complex nature of SSR loci, however, demonstrates the critical value of sequence level analysis for phylogenetic inference.

The flanking regions of gene-associated SSRs were highly conserved within and between endophyte taxa (80%–100% sequence identity across coding and non-coding regions), supporting a common origin for these species [1, 9]. Despite this level of sequence conservation, SSR-flanking regions were informative for studying genetic relationships. The different individual loci obtained similar genetic relationships, consistent with previous studies of other genes. Differences in the power of the individual loci to resolve relationships were identified due to variation in number of informative characters and composition (exon, intron, coding, or non-coding) of amplicons. In other studies, SSR-flanking sequences from different loci have been aggregated to increase the number of informative characters and improve resolution of phylogenetic relationships [31, 32, 36, 37]. However, variation of inferred ploidy level and of number of haplotypes derived from different loci in heteroploids would potentially bias such aggregation studies. As a consequence, each locus was analysed separately in this study.

4.2. Genome Affinities between Neotyphodium and Epichloë Species. The close relationships between taxa were in accordance with those predicted from genes more commonly used for phylogenetic analysis such as rDNA, *tubB*, *tefA,* and *actG*. Single locus-specific haplotypes were obtained from all *Epichloë* species and from *N. lolii*, the latter being closely related to *E. festucae*. Other *Neotyphodium* species contained multiple haplotypes that were similar to those from different *Epichloë* species. Occurrence of different haplotype subclasses in *N. coenophialum*, FaTG-2, and *N. uncinatum* is consistent with the heteroploid or nonhaploid nature of these species [11, 14] and indicates the presence of multiple genomes (Figure 6).

4.3. Relationships with Unclassified Neotyphodium Isolates. Neotyphodium isolates that could not be assigned to known morphological classes also appear to differ from characterised taxa at the molecular level [17]. Isolate 9303/2 and isolate 9728 have been assigned to taxonomic groupings HeuTG-2 and LpTG-2, respectively, based on phylogenetic analysis of the *tefA* and *tubB* genes (A. Leuchtmann, pers. comm.; [17]). Analysis of SSR-flanking regions in this study, however, suggests that the isolates show closer affinities than formerly predicted. Moon et al. [17] reported the detection of two haplotype classes for each isolate, closely related to those from *E. bromicola* and *E. typhina* (HeuTG) and from *E. festucae* and *E. typhina* (LpTG-2). These phylogenetic affinities were also detected in the current study. However,

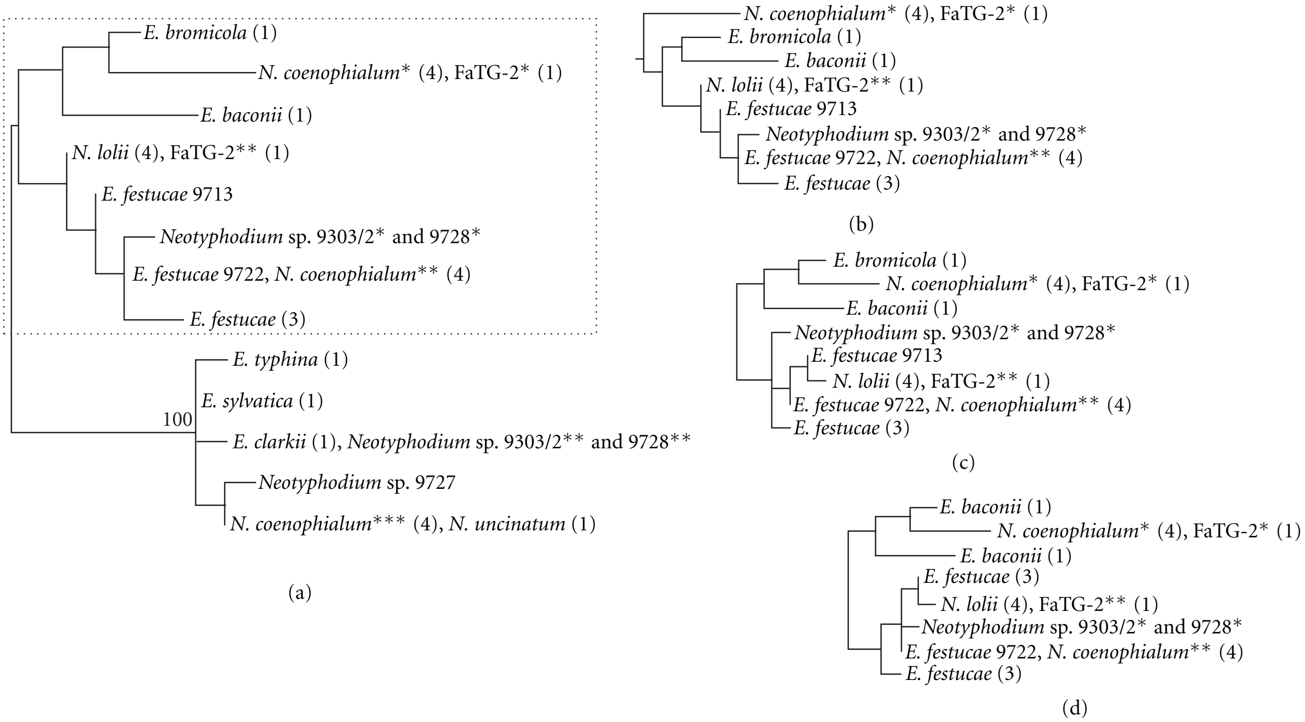

E. bromicola (1)
N. coenophialum* (4), FaTG-2* (1)
E. baconii (1)
N. lolii (4), FaTG-2** (1)
E. festucae 9713
Neotyphodium sp. 9303/2* and 9728*
E. festucae 9722, N. coenophialum** (4)
E. festucae (3)
E. typhina (1)
E. sylvatica (1)
100
E. clarkii (1), Neotyphodium sp. 9303/2** and 9728**
Neotyphodium sp. 9727
N. coenophialum*** (4), N. uncinatum (1)
(a)
N. coenophialum* (4), FaTG-2* (1)
E. bromicola (1)
E. baconii (1)
N. lolii (4), FaTG-2** (1)
E. festucae 9713
Neotyphodium sp. 9303/2* and 9728*
E. festucae 9722, N. coenophialum** (4)
E. festucae (3)
(b)
E. bromicola (1)
N. coenophialum* (4), FaTG-2* (1)
E. baconii (1)
Neotyphodium sp. 9303/2* and 9728*
E. festucae 9713
N. lolii (4), FaTG-2** (1)
E. festucae 9722, N. coenophialum** (4)
E. festucae (3)
(c)
E. baconii (1)
N. coenophialum* (4), FaTG-2* (1)
E. baconii (1)
E. festucae (3)
N. lolii (4), FaTG-2** (1)
Neotyphodium sp. 9303/2* and 9728*
E. festucae 9722, N. coenophialum** (4)
E. festucae (3)
(d)

Figure 3: Parsimony analysis of sequence haplotypes derived from reference *Neotyphodium* and *Epichloë* isolates for the EST-SSR locus NCESTA1GA07. (a) One of four most parsimonious trees found. Diagram properties are as described in the legend to Figure 1 ((b), (c), and (d)) Subtrees showing the alternative topologies of the *E. festucae*, *E. baconii,* and *E. bromicola* clade in the other trees.

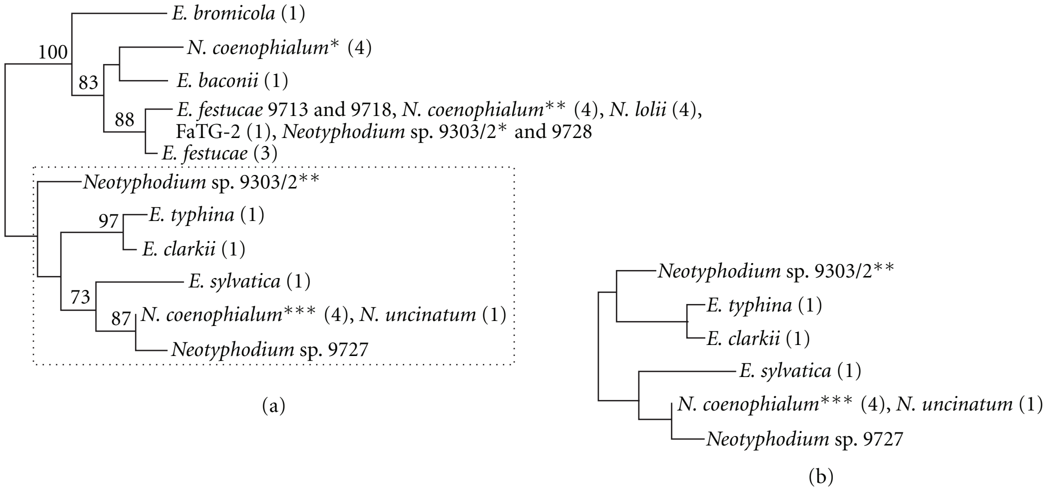

E. bromicola (1)
100
N. coenophialum* (4)
83
E. baconii (1)
88
E. festucae 9713 and 9718, N. coenophialum** (4), N. lolii (4), FaTG-2 (1), Neotyphodium sp. 9303/2* and 9728
E. festucae (3)
Neotyphodium sp. 9303/2**
97
E. typhina (1)
E. clarkii (1)
E. sylvatica (1)
73
87
N. coenophialum*** (4), N. uncinatum (1)
Neotyphodium sp. 9727
(a)
Neotyphodium sp. 9303/2**
E. typhina (1)
E. clarkii (1)
E. sylvatica (1)
N. coenophialum*** (4), N. uncinatum (1)
Neotyphodium sp. 9727
(b)

Figure 4: Parsimony analysis of sequence haplotypes derived from reference *Neotyphodium* and *Epichloë* isolates for the EST-SSR locus NLESTA1GF09. (a) One of two most parsimonious trees found. Diagram properties are as described in the legend to Figure 1. (b) Subtree showing the alternative topology of the *E. typhina*, *E. clarkii,* and *E. sylvatica* clade in the other tree.

both flanking sequence analysis, as well as phenetic studies based on a larger number of SSR loci [38], detected a third haplotype sub-class for both isolates and suggested common affinities with *E. festucae*, *E. bromicola*, and *E. typhina*, respectively. Accurate inference of phylogenetic relationships among *Neotyphodium* and *Epichloë* species consequently requires characterisation of a number of different genomic loci.

Although isolates 9303/2 and 9728 share common affinities, DNA-based phylogenetic and phenetic analyses suggest mutual genetic divergence and placement in taxonomic groups with different relative gene content. Although similar-sized haplotypes were detected, SSR polymorphism between these isolates was greater than that detected within *N. coenophialum*, *N. lolii,* and *E. festucae* [38]. In addition,

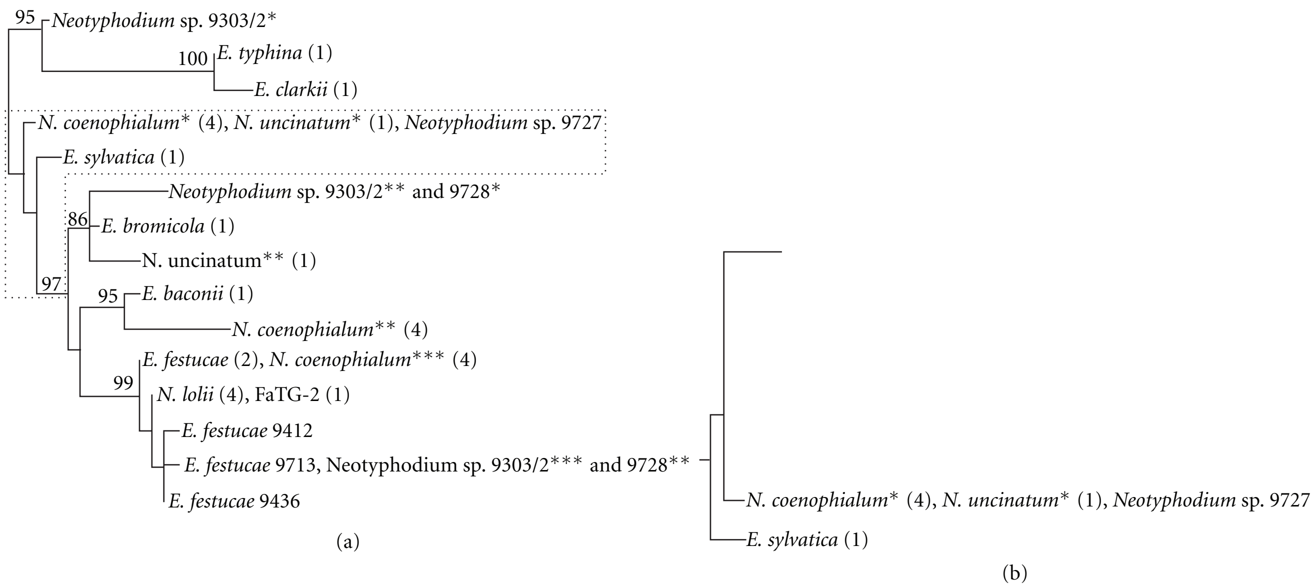

95
Neotyphodium sp. 9303/2*
100
E. typhina (1)
E. clarkii (1)
N. coenophialum* (4), N. uncinatum* (1), Neotyphodium sp. 9727
E. sylvatica (1)
Neotyphodium sp. 9303/2** and 9728*
86
E. bromicola (1)
N. uncinatum** (1)
97
95
E. baconii (1)
N. coenophialum** (4)
E. festucae (2), N. coenophialum*** (4)
99
N. lolii (4), FaTG-2 (1)
E. festucae 9412
E. festucae 9713, Neotyphodium sp. 9303/2*** and 9728**
E. festucae 9436
(a)
N. coenophialum* (4), N. uncinatum* (1), Neotyphodium sp. 9727
E. sylvatica (1)
(b)

Figure 5: Parsimony analysis of sequence haplotypes derived from reference *Neotyphodium* and *Epichloë* isolates for the EST-SSR locus NLESTA1NF04. (a) One of two most parsimonious trees found. Diagram properties are as described in the legend to Figure 1. (b) Sub-tree showing the alternative branching of *E. sylvatica* in the other tree. Note that the nucleotide sequence of the second amplification product from FaTG-2 and the third amplification product from unidentified *Neotyphodium* isolate 9728 detected at this locus by autoradiography [38] was not obtained.

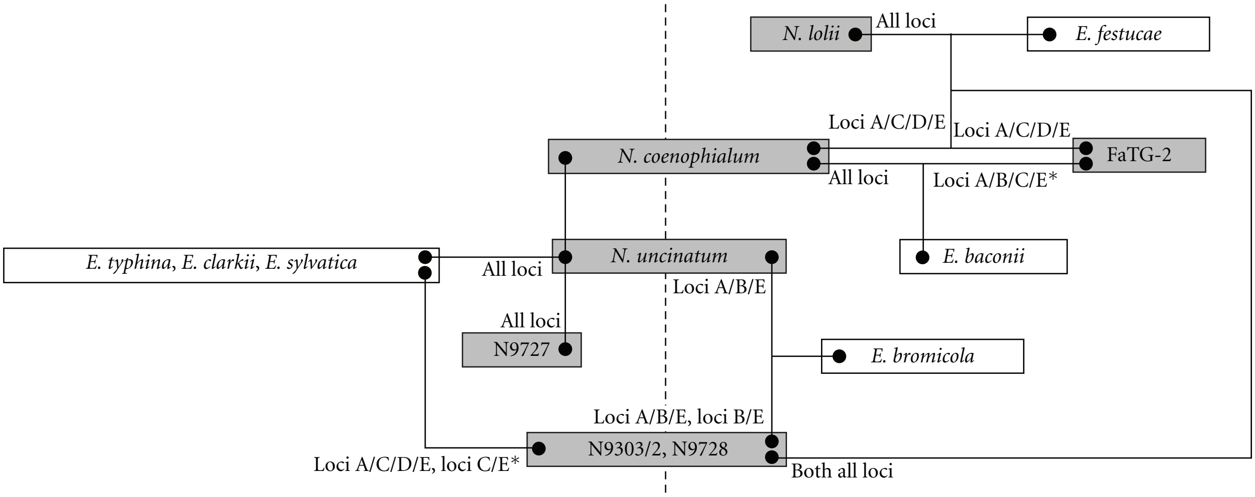


Figure 6: Summary of genomic affinities between *Neotyphodium* and *Epichloë* species predicted from phylogenetic analysis of the flanking regions of five SSR loci: NCESTA1AB04 (a), NCESTA1FH03 (b), NCESTA1GA07 (c), NLESTA1GF09 (d), and NLESTA1NF04 (e). Predicted partial or complete genomes, based on variant haplotype sub-classes, are indicated as black circles. Lines that connect putative common genomes, and the level of support for each inference are indicated in terms of the number of loci providing confirmatory data. This information relates to the next most adjacent taxon in the topology of the diagram. An asterisk indicates a locus-specific variant obtained by PCR, but for which the nucleotide sequence was not obtained. The dotted line defines the inferred division between the two major groups of *Epichloë* species.

9303/2 and 9728 failed to cluster together in an AFLP-derived phenogram [38], which represents a genome-wide assessment of genetic polymorphism. SSR polymorphism analysis also detected substantial differences in the number of locus-specific haplotypes: isolate 9728 produced a higher proportion of single haplotype classes. DNA sequence analysis further demonstrated differences in both number and type of haplotype. *E. festucae*-related haplotypes were detected in both isolates for all five loci, while *E. bromicola*-like and *E. typhina*-like sequence variants were observed more frequently in isolate 9303/2 than isolate 9728 and were not represented between all loci. *Neotyphodium* species with common phylogenetic affinities are known to occur in several different grass species. The LpTG-2, *N. tembladerae,* and *N. australiense* endophytes, which are resident in *L. perenne*, *Poa huecu,* and *Echinopogon ovatus*, respectively, all appear to be phylogenetically related to *E. festucae* and *E. typhina* [10, 16]. However, these endophytes appear to be related to different *E. typhina* strains and also differ in their genome structure [10, 13, 16]. Differences in transcript levels associated with different gene-specific sequence variants, as observed for the 60S ribosomal protein-encoding gene in this study, may also contribute to phenotypic trait variation between different heteroploid endophyte taxa.

Two asexual endophyte species, *N. huerfanum* and *N. tembladerae*, are known to occur in *Festuca arizonica* [17]. Phylogenetic analysis of the third unclassified *Neotyphodium* isolate (9727), which was also derived from *F. arizonica*, suggests that it may belong to the former taxon. Isolate 9727 produced single haplotypes and these sequences, like those of the *N. huerfanum tefA* and *tubB* loci [17], are closely related to the inferred *E. typhina*-related haplotypes from *N. coenophialum* and *N. uncinatum* (Section 4.2). These results were also supported by SSR polymorphism-based phenetic analysis, in which 9727 clustered with *E. typhina*, *E. clarkii,* and *E. sylvatica*, while in AFLP analysis the isolate clustered with *N. uncinatum* [38].

4.4. Origins of Neotyphodium and Epichloë Species. Due to their close phylogenetic relationships with specific *Epichloë* species, *Neotyphodium* species have been proposed to have originated from these sexual endophyte taxa either directly through the loss of the sexual state, or through interspecific hybridisation of distinct *Epichloë* and *Neotyphodium* species. The first process is proposed to have given rise to haploid *Neotyphodium* species such as *N. lolii*, while the heteroploid *Neotyphodium* species such as *N. coenophialum*, FaTG-2, and *N. uncinatum* may have arisen through the second evolutionary process. Because *Epichloë* species form unique mating populations [49–51] and *Neotyphodium* species are not known to sporulate *in vivo* [52], this second mode of evolution is thought to have been a parasexual process involving somatic fusion of endophyte hyphae. This hypothesis does, however, require physical colocation between endophyte taxa that generally occur in distinct host species. In addition, mechanisms of gene loss following nuclear fusion are necessary to account for the observed genomic composition of contemporary heteroploid taxa, as a range of studies [2, 10, 11, 13, 14, 16, 17] have shown that extant *Neotyphodium* species do not appear to have the full complement of genes present in phylogenetically related *Epichloë* species. Loss of genes involved in sexual reproduction and pathogenicity would be a prerequisite for such genomic rearrangement events, as well as genes vulnerable to dosage-dependent effects. It is also formally possible that sexual *Epichloë* species may have arisen from asexual *Neotyphodium* species in response to selective environmental pressures, a mechanism requiring both gene loss and gene gain, possibly through horizontal gene transfer. Mechanisms for both processes have been inferred through comparisons of different fungal taxa at the whole genome levels [53] and may have been facilitated by structural features such as presence of conserved repetitive elements.

In conclusion, this study demonstrates the application of SSR-flanking sequences to studies of genome affinities between pasture grass fungal endophyte species for clarification of novel modes of genome evolution. The inferred affinities were consistent with those obtained from gene loci that are more commonly used in molecular phylogenetics, but provided a more extensive survey of genomic loci, that may be ultimately extended to whole genome comparisons based.on second-generation sequencing technologies.

Acknowledgments

Endophytes isolates were kindly provided by Dr. Adrian Leuchtmann (ETH Zürich, Switzerland) and Nola McFarlane (DPI-Hamilton). This work was supported by the Victorian Department of Primary Industries and the Molecular Plant Breeding Cooperative Research Centre. All experiments conducted during this study comply with current Australian laws.

References

[1] G. A. Kuldau, J. S. Liu, J. F. White, M. R. Siegel, and C. L. Schardl, "Molecular systematics of Clavicipitaceae supporting monophyly of genus *Epichloë* and form genus Ephelis," *Mycologia*, vol. 89, no. 3, pp. 431–441, 1997.

[2] K. Clay and C. Schardl, "Evolutionary origins and ecological consequences of endophyte symbiosis with grasses," *American Naturalist*, vol. 160, supplement S, pp. S99–S127, 2002.

[3] R. T. Gallagher, E. P. White, and P. H. Mortimer, "Ryegrass staggers: isolation of potent neurotoxins lolitrem A and lolitrem B from staggers-producing pastures," *New Zealand Veterinary Journal*, vol. 29, no. 10, pp. 189–190, 1981.

[4] S. G. Yates, R. D. Plattner, and G. B. Garner, "Detection of ergopeptine alkaloids in endophyte infected, toxic Ky-31 tall fescue by mass spectrometry/mass spectrometry," *Journal of Agricultural and Food Chemistry*, vol. 33, no. 4, pp. 719–722, 1985.

[5] D. D. Rowan and D. L. Gaynor, "Isolation of feeding deterrents against argentine stem weevil from ryegrass infected with the endophyte *Acremonium loliae*," *Journal of Chemical Ecology*, vol. 12, no. 3, pp. 647–658, 1986.

[6] M. Arachevaleta, C. W. Bacon, C. S. Hoveland et al., "Effect of the tall fescue endophyte on plant response to environmental stress," *Agronomy Journal*, vol. 81, no. 1, pp. 83–90, 1989.

[7] C. Ravel, C. Courty, A. Coudret, and G. Charmet, "Beneficial effects of Neotyphodium lolii on the growth and the water status in perennial ryegrass cultivated under nitrogen deficiency or drought stress," *Agronomie*, vol. 17, no. 3, pp. 173–181, 1997.

[8] G. Morgan-Jones and W. Gams, "Notes on Hyphomycetes. XLI. An endophytes of *Festuca arundinacea* and the anamorph of *Epichloë typhina*, new taxa in one of two new sections of Acremonium," *Mycotaxon*, vol. 15, no. 1, pp. 311–318, 1982.

[9] A. E. Glenn, C. W. Bacon, R. Price, and R. T. Hanlin, "Molecular phylogeny of *Acremonium* and its taxonomic implications," *Mycologia*, vol. 88, no. 3, pp. 369–383, 1996.

[10] C. L. Schardl, A. Leuchtmann, H. F. Tsai, M. A. Collett, D. M. Watt, and D. B. Scott, "Origin of a fungal symbiont of perennial ryegrass by interspecific hybridization of a mutualist with the ryegrass choke pathogen, *Epichloë typhina*," *Genetics*, vol. 136, no. 4, pp. 1307–1317, 1994.

[11] H. F. Tsai, J. S. Liu, C. Staben et al., "Evolutionary diversification of fungal endophytes of tall fescue grass by hybridization with *Epichloë* species," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 91, no. 7, pp. 2542–2546, 1994.

[12] C. L. Schardl, A. Leuchtmann, K. R. Chung, D. Penny, and M. R. Siegel, "Coevolution by common descent of fungal symbionts (*Epichloë* spp.) and grass hosts," *Molecular Biology and Evolution*, vol. 14, no. 2, pp. 133–143, 1997.

[13] C. D. Moon, B. Scott, C. L. Schardl, and M. J. Christensen, "The evolutionary origins of *Epichloë* endophytes from annual ryegrasses," *Mycologia*, vol. 92, no. 1–6, pp. 1103–1118, 2000.

[14] K. D. Craven, J. D. Blankenship, A. Leuchtmann, K. Hignight, and C. L. Schardl, "Hybrid fungal endophytes symbiotic with the grass *Lolium pratense*," *Sydowia*, vol. 53, no. 1, pp. 44–73, 2001.

[15] K. D. Craven, P. T. W. Hsiau, A. Leuchtmann, W. Hollin, and C. L. Schardl, "Multigene phylogeny of *Epichloë* species, fungal symbionts of grasses," *Annals of the Missouri Botanical Garden*, vol. 88, no. 1, pp. 14–34, 2001.

[16] C. D. Moon, C. O. Miles, U. Järlfors, and C. L. Schardl, "The evolutionary origins of three new *Neotyphodium* endophyte species from grasses indigenous to the Southern Hemisphere," *Mycologia*, vol. 94, no. 4, pp. 694–711, 2002.

[17] C. D. Moon, K. D. Craven, A. Leuchtmann, S. L. Clement, and C. L. Schardl, "Prevalence of interspecific hybrids amongst asexual fungal endophytes of grasses," *Molecular Ecology*, vol. 13, no. 6, pp. 1455–1467, 2004.

[18] G. A. Kuldau, H. F. Tsai, and C. L. Schardl, "Genome sizes of *Epichloë* species and anamorphic hybrids," *Mycologia*, vol. 91, no. 5, pp. 776–782, 1999.

[19] C. L. Schardl, J. S. Liu, J. F. White, R. A. Finkel, Z. An, and M. R. Siegel, "Molecular phylogenetic relationships of nonpathogenic grass mycosymbionts and clavicipitaceous plant pathogens," *Plant Systematics and Evolution*, vol. 178, no. 1-2, pp. 27–41, 1991.

[20] D. Tautz, "Hypervariability of simple sequences as a general source for polymorphic DNA markers," *Nucleic Acids Research*, vol. 17, no. 16, pp. 6463–6471, 1989.

[21] J. L. Weber and P. E. May, "Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction," *American Journal of Human Genetics*, vol. 44, no. 3, pp. 388–396, 1989.

[22] M. D. Schug, C. M. Hutter, K. A. Wetterstrand, M. S. Gaudette, T. F. C. Mackay, and C. F. Aquadro, "The mutation rates of di-, tri- and tetranucleotide repeats in *Drosophila melanogaster*," *Molecular Biology and Evolution*, vol. 15, no. 12, pp. 1751–1760, 1998.

[23] G. Levinson and G. A. Gutman, "Slipped-strand mispairing: a major mechanism for DNA sequence evolution," *Molecular Biology and Evolution*, vol. 4, no. 3, pp. 203–221, 1987.

[24] O. Rose and D. Falush, "A threshold size for microsatellite expansion," *Molecular biology and evolution*, vol. 15, no. 5, pp. 613–615, 1998.

[25] J. C. Garza, M. Slatkin, and N. B. Freimer, "Microsatellite allele frequencies in humans and chimpanzees, with implications for constraints on allele size," *Molecular Biology and Evolution*, vol. 12, no. 4, pp. 594–603, 1995.

[26] D. B. Goldstein and A. G. Clark, "Microsatellite variation in North American populations of *Drosophila melanogaster*," *Nucleic Acids Research*, vol. 23, no. 19, pp. 3882–3886, 1995.

[27] M. C. Grimaldi and B. Crouau-Roy, "Microsatellite allelic homoplasy due to variable flanking sequences," *Journal of Molecular Evolution*, vol. 44, no. 3, pp. 336–340, 1997.

[28] I. Colson and D. B. Goldstein, "Evidence for complex mutations at microsatellite loci in Drosophila," *Genetics*, vol. 152, no. 2, pp. 617–627, 1999.

[29] I. Clisson, M. Lathuilliere, and B. Crouau-Roy, "Conservation and evolution of microsatellite loci in primate taxa," *American Journal of Primatology*, vol. 50, no. 3, pp. 205–214, 2000.

[30] X. Chen, Y. G. Cho, and S. R. McCouch, "Sequence divergence of rice microsatellites in *Oryza* and other plant species," *Molecular Genetics and Genomics*, vol. 268, no. 3, pp. 331–343, 2002.

[31] S. R. Santos, T. L. Shearer, A. R. Hannes, and M. A. Coffroth, "Fine-scale diversity and specificity in the most prevalent lineage of symbiotic dinoflagellates (*Symbiodinium*, Dinophyceae) of the Caribbean," *Molecular Ecology*, vol. 13, no. 2, pp. 459–469, 2004.

[32] T. Asahida, A. K. Gray, and A. J. Gharrett, "Use of microsatellite locus flanking regions for phylogenetic analysis? A preliminary study of *Sebastes* subgenera," *Environmental Biology of Fishes*, vol. 69, no. 1-4, pp. 461–470, 2004.

[33] D. Field and C. Wills, "Abundant microsatellite polymorphism in *Saccharomyces cerevisiae*, and the different distributions of microsatellites in eight prokaryotes and *S. cerevisiae*, result from strong mutation pressures and a variety of selective forces," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 95, no. 4, pp. 1647–1652, 1998.

[34] M. S. Röder, V. Korzun, K. Wendehake et al., "A microsatellite map of wheat," *Genetics*, vol. 149, no. 4, pp. 2007–2023, 1998.

[35] M. Morgante, M. Hanafey, and W. Powell, "Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes," *Nature Genetics*, vol. 30, no. 2, pp. 194–200, 2002.

[36] M. C. Fisher, G. Koenig, T. J. White, and J. W. Taylor, "A test for concordance between the multilocus genealogies of genes and microsatellites in the pathogenic fungus *Coccidioides immitis*," *Molecular Biology and Evolution*, vol. 17, no. 8, pp. 1164–1174, 2000.

[37] M. Rossetto, J. McNally, and R. J. Henry, "Evaluating the potential of SSR flanking regions for examining taxonomic relationships in the Vitaceae," *Theoretical and Applied Genetics*, vol. 104, no. 1, pp. 61–66, 2002.

[38] E. Van Zijll De Jong, K. M. Guthridge, G. C. Spangenberg, and J. W. Forster, "Development and characterization of EST-derived simple sequence repeat (SSR) markers for pasture grass endophytes," *Genome*, vol. 46, no. 2, pp. 277–290, 2003.

[39] S. F. Altschul, T. L. Madden, A. A. Schäffer et al., "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs," *Nucleic Acids Research*, vol. 25, no. 17, pp. 3389–3402, 1997.

[40] J. D. Thompson, T. J. Gibson, F. Plewniak, F. Jeanmougin, and D. G. Higgins, "The CLUSTAL X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools," *Nucleic Acids Research*, vol. 25, no. 24, pp. 4876–4882, 1997.

[41] A. R. Templeton, "Phylogenetic inference from restriction endonuclease cleavage site maps with particular reference to the evolution of humans and the apes," *Evolution*, vol. 37, no. 2, pp. 221–244, 1983.

[42] H. Kishino and M. Hasegawa, "Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from

DNA sequence data, and the branching order in hominoide," *Journal of Molecular Evolution*, vol. 29, no. 2, pp. 170–179, 1989.

[43] H. Shimodaira and M. Hasegawa, "Multiple comparisons of log-likelihoods with applications to phylogenetic inference," *Molecular Biology and Evolution*, vol. 16, no. 8, pp. 1114–1116, 1999.

[44] J. Felsenstein, "Confidence limits on phylogenies: an approach using the bootstrap," *Evolution*, vol. 39, no. 4, pp. 783–791, 1985.

[45] J. Felsenstein and G. A. Churchill, "A Hidden Markov Model approach to variation among sites in rate of evolution," *Molecular Biology and Evolution*, vol. 13, no. 1, pp. 93–104, 1996.

[46] W. M. Fitch and E. Margoliash, "Construction of phylogenetic trees," *Science*, vol. 155, no. 760, pp. 279–284, 1967.

[47] N. Saitou and M. Nei, "The neighbor-joining method: a new method for reconstructing phylogenetic trees," *Molecular biology and evolution*, vol. 4, no. 4, pp. 406–425, 1987.

[48] H. A. Schmidt, K. Strimmer, M. Vingron, and A. Von Haeseler, "TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing," *Bioinformatics*, vol. 18, no. 3, pp. 502–504, 2002.

[49] A. Leuchtmann, C. L. Schardl, and M. R. Siegel, "Sexual compatibility and taxonomy of a new species of *Epichloë* symbiotic with fine fescue grasses," *Mycologia*, vol. 86, no. 6, pp. 802–812, 1994.

[50] A. Leuchtmann and C. L. Schardl, "Mating compatibility and phylogenetic relationships among two new species of *Epichloë* and other congeneric European species," *Mycological Research*, vol. 102, no. 10, pp. 1169–1182, 1998.

[51] C. L. Schardl and A. Leuchtmann, "Three new species of *Epichloë* symbiotic with North American grasses," *Mycologia*, vol. 91, no. 1, pp. 95–107, 1999.

[52] J. F. White, "Endophyte-host associations in forage grasses. XI. A proposal concerning origin and evolution," *Mycologia*, vol. 80, no. 4, pp. 442–446, 1988.

[53] E. L. Braun, A. L. Halpern, M. A. Nelson, and D. O. Natvig, "Large-scale comparison of fungal sequence information: mechanisms of innovation in Neurospora crassa and gene loss in Saccharomyces cerevisiae," *Genome Research*, vol. 10, no. 4, pp. 416–430, 2000.

SAGE-Hindawi Access to Research
International Journal of Evolutionary Biology
Volume 2011, Article ID 423821, 7 pages
doi:10.4061/2011/423821

Research Article

Evolutionary Origins of the Fumonisin Secondary Metabolite Gene Cluster in *Fusarium verticillioides* and *Aspergillus niger*

Nora Khaldi[1] and Kenneth H. Wolfe[2]

[1] *UCD Conway Institute of Biomolecular and Biomedical Research, UCD School of Medicine and Medical Sciences, and UCD Complex and Adaptive Systems Laboratory, University College Dublin, Dublin 4, Ireland*
[2] *Smurfit Institute of Genetics, Trinity College Dublin, Dublin 2, Ireland*

Correspondence should be addressed to Nora Khaldi, khaldin@tcd.ie

Received 15 October 2010; Revised 10 January 2011; Accepted 15 March 2011

Academic Editor: Hiromi Nishida

The secondary metabolite gene clusters of euascomycete fungi are among the largest known clusters of functionally related genes in eukaryotes. Most of these clusters are species specific or genus specific, and little is known about how they are formed during evolution. We used a comparative genomics approach to study the evolutionary origins of a secondary metabolite cluster that synthesizes a polyketide derivative, namely, the fumonisin (FUM) cluster of *Fusarium verticillioides*, and that of *Aspergillus niger* another fumonisin (fumonisin B) producing species. We identified homologs in other euascomycetes of the *Fusarium verticillioides FUM* genes and their flanking genes. We discuss four models for the origin of the *FUM* cluster in *Fusarium verticillioides* and argue that two of these are plausible: (i) assembly by relocation of initially scattered genes in a recent *Fusarium verticillioides*; or (ii) horizontal transfer of the *FUM* cluster from a distantly related Sordariomycete species. We also propose that the *FUM* cluster was horizontally transferred into *Aspergillus niger*, most probably from a Sordariomycete species.

1. Introduction

The order of genes along eukaryotic chromosomes is often assumed to be random, but there is growing evidence that the chromosomal position of some genes is maintained by natural selection [1–3], and that selection can sometimes operate to move genes to new locations [4–6]. Among the eukaryotes, many of the most notable examples of the physical clustering of genes with related functions occur in the filamentous fungi [7–10]. The most striking fungal gene clusters are those involved in the synthesis of secondary metabolites such as sterigmatocystin (a 25-gene cluster in *Aspergillus nidulans* [11]), fumonisin (a 17-gene cluster in *Fusarium verticillioides* [12, 13]), and trichothecene (a 12-gene cluster in *Fusarium sporotrichioides* [14]). Secondary metabolites are organic molecules that are not essential for the normal growth of the fungus but which function in host/pathogen interactions or other forms of communication or warfare between organisms. They are typically modified polyketides, terpenes, alkaloids, or nonribosomal peptides [15]. Synthesis of these molecules requires many successive enzymatic steps, and the genes for these enzymes are almost invariably clustered together at a single genomic location. It has been suggested that physical clustering may allow the genes to be coregulated by means of chromatin modification [15], or that it may facilitate horizontal transfer of intact clusters between species [16].

Since the discovery of secondary metabolite gene clusters, their mechanism of origin and assembly has remained a matter of speculation. The growing number of available euascomycete genome sequences now enables us to both predict new secondary metabolite clusters [17] and take a phylogenomic approach to the evolutionary origins of these clusters. In this study we focus on one of the largest known secondary metabolite clusters “Fumonisin”, a polyketide synthase type cluster. Fumonisins (FB/FC) are mycotoxins produced by some species in the *Gibberella fujikuroi* species complex, of which *F. verticillioides* and *F. proliferatum* are the most studied (*Gibberella* is a teleomorphic form of the genus *Fusarium*). The *FUM* genes are organized as a cluster of 17 genes in both species (including *FUM20* and *FUM21*, two additional genes recently revealed in the cluster [13]), though the location of the cluster differs between the two [12, 18]. The products of the *FUM* cluster genes include a polyketide synthase, fatty acyl-CoA synthases, and cytochrome P450 monooxygenases. Expression of the genes in the cluster, but not the neighboring genes, is induced under conditions when fumonisin (FB/FC) is synthesized [12]. The ability to synthesize fumonisin (FB/FC) has a patchy phylogenetic distribution across the genus *Fusarium*, due to the variable presence or absence of the *FUM* cluster among different isolates [19]. *F. graminearum*, for instance, does not synthesize fumonisin.

Recently, fumonisin (FB) production has also been reported in *Aspergillus niger* [20]. The genes in this species are also clustered and homologous to the *FUM* genes of *F. verticillioides* [21]. This is surprising given the large evolutionary distance between *A. niger* and *F. verticillioides*.

We combined comparative genomics with phylogenetic analysis to investigate whether genes in the *FUM* cluster have homologs in filamentous fungi that do not synthesize the mycotoxins, and so to study how this cluster was formed.

2. Methods

Our analysis was done using the completely sequenced genomes of the euascomycetes *A. nidulans* [22], *F. graminearum* [23], *M. grisea* [24], and *N. crassa* [25]. A set of 87,000 expressed sequence tags from *F. verticillioides* [26] was used for analysis of rates of sequence evolution in this species compared to *F. graminearum*.

To identify homologs of genes in the *FUM* clusters, we first used each protein as a query in a BLASTP search against the NCBI nonredundant protein sequence database. Because the *F. verticillioides* genome is not present in this database, we made a local BLAST database for expressed sequence tags from this species. Sequences giving hits with Expect (E) values of less than $1e - 4$ were retained and used for phylogenetic analysis. Each set of proteins was aligned using ClustalW [27] and poorly aligned regions were removed using Gblocks [28]. Maximum likelihood trees were constructed using PHYML [29] with the JTT amino acid substitution matrix and four categories of substitution rates. Bootstrapping was done using the default options in PHYML with 100 replicates per run. The trees were eyeballed, the distant sequences were removed, and the steps above (ClustalW, Gblocks, and PHYML) were repeated on the remaining sequences.

3. Results

In the following we are interested in identifying scenarios that can explain the origins of the current *FUM* gene cluster. We do this by listing all possible scenarios that might have taken place in evolution, and comparing their plausibility. Building an evolutionary scenario is not straightforward because many of the events took place in extinct species and only a few clues remain in the current organisms. Almost any scenario is formally possible, but what makes one scenario more likely than another is parsimony-consideration of the number of separate events that are required to have taken place in order to account for it. In other words, a scenario involving fewer events is a more likely explanation of the observed data.

3.1. Origins of the FUM Gene Cluster in F. verticillioides.

Gene-by-gene phylogenetic analyses were carried out to decipher the evolutionary history of the *FUM* genes using homologs of these genes in four euascomycetes: *Fusarium graminearum*, *Neurospora crassa*, *Magnaporthe grisea*, and *Aspergillus nidulans* (we use the word “homologs” for convenience in situations where we are unsure whether genes are orthologs or paralogs). We also constructed individual phylogenies for the genes located on either side of the *F. verticillioides FUM* cluster. Homology relationships between the genes in or near the *F. verticillioides FUM* cluster and other euascomycete genes are summarized in Figure 1. We identified probable orthologs of 13 of the 17 *FUM* genes (Figure 1). These genes are not arranged in clusters in the other genomes.

A region of *F. graminearum* chromosome 1 (genes *FG00269–F00276*; Figure 1) contains orthologs of the five genes to the left of the *FUM* cluster (*F. verticillioides NPT1*, *WDR1*, *PNG1*, *ZNF1*, and *ZBD1*) immediately adjacent to orthologs of the two genes to the right of the cluster (*F. verticillioides ORF21* and *MPU1*), with nothing in between them in *F. graminearum*. Similarly, in *M. grisea* the *ZBD1* ortholog is beside the *ORF21* ortholog, and in *A. nidulans* the *PNG1* ortholog is beside the *ORF21* ortholog. Thus, chromosomal sites orthologous to the *F. verticillioides FUM* cluster-flanking regions exist adjacent to one another in *F. graminearum*, *M. grisea,* and *A. nidulans*, but no orthologs of the *FUM* genes themselves are found at these sites (Figure 1). Instead, homologs of the *FUM* genes are scattered on different chromosomes of the genomes of these fumonisin-nonproducing species. In *F. graminearum*, for instance, the 10 homologs of *FUM* genes are dispersed across four chromosomes and none of them is located close to another or to the *FUM*-flanking genes. Thus the *FUM* cluster genes appear to have been inserted into a pre-existing genomic locus between *ZBD1* and *ORF21*.

Further, we examined the genomic contexts around each of the *FUM* cluster homologs in other species. For example, the *F. verticillioides FUM* cluster gene *FUM11* is homologous to *FG07875* in *F. graminearum* and to *MG03479* in *M. grisea* (Figure 1). We will refer to *FG07875* and *MG03479* as focal genes. When we examine the regions around these focal genes in the *F. graminearum* and *M. grisea* genomes, we find that some of the neighboring genes near them are also orthologs of each other. On one side *FG07874* (1 gene away from the focal gene) is an ortholog of *MG03474* (5 genes away), and on the other side *FG07864* (11 genes away from the focal gene) is an ortholog of *MG03480* (1 gene away from the focal gene). In Figure 1, only the focal genes are shown but these similarities of context are indicated by the

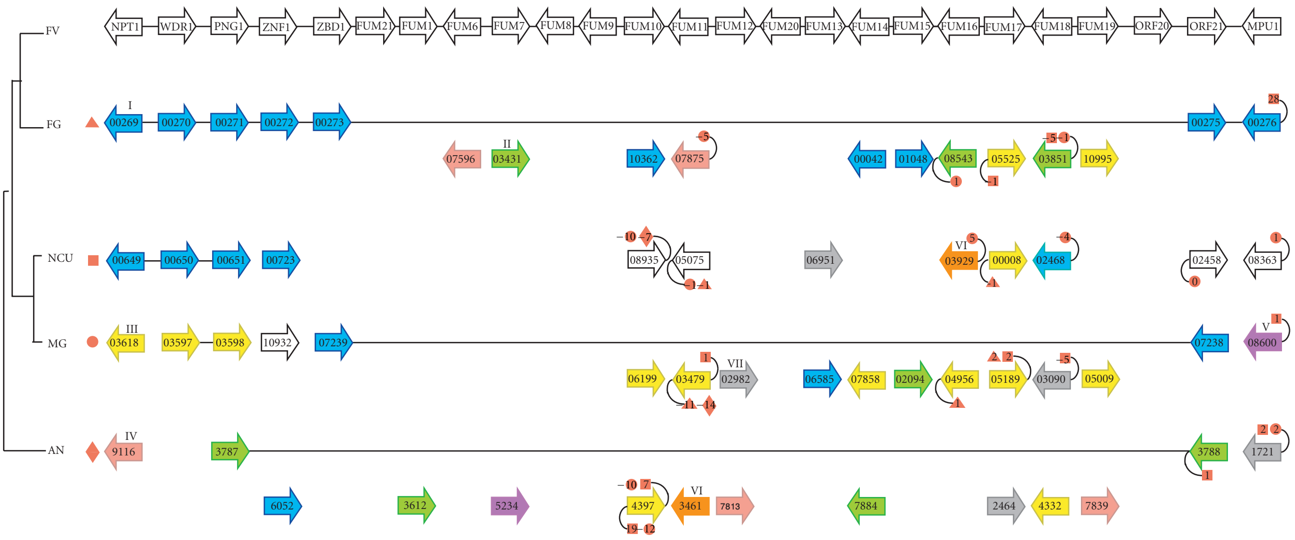


Figure 1: Representation of the fumonisin cluster and its flanking genes in *F. verticillioides* (FV). Columns are homologs of the FUM genes and orthologs of the flanking genes identified by phylogenetic analysis in *F. graminearum* (FG), *N. crassa* (NCU), *M. grisea* (MG), and *A. nidulans* (AN). Genes in the latter four species are identified by gene numbers from their genome projects. Different colors represent different chromosomes. The long lines in F. graminearum, M. grisea, and A. nidulans show that in those species, there is a site in the genome that corresponds to the FUM cluster location, but no FUM genes are present at that locus. Curved lines and numbers in orange symbols indicate conservation of the neighboring genes around FUM homologs in F. graminearum, N. crassa, M. grisea, and A. nidulans. For each gene show in the figure (the focal genes) we considered the two genes immediately next to it. If these genes have orthologs located <20 genes away from the focal gene's ortholog in another species, a symbol indicates this fact. For example, the numbers −10 (in a circle) and −7 (in a diamond) connected to gene NCU08935 indicate that the gene immediately after NCU08935 (i.e., NCU08936) has an ortholog in M. grisea that is 10 genes away from MG06199 (i.e., MG06189), and an ortholog in A. nidulans that is 7 genes away from AN04397 (i.e., AN04392). Triangles, squares, circles, and diamonds indicate relationships to *F. graminearum*, *N. crassa*, *M. grisea,* and *A. nidulans*, respectively.

small numbers at the ends of the curved lines attached to the symbols for the focal genes *FG07875* and *MG03479*.

Overall, the homologs of five *FUM* cluster genes (*FUM10*, *FUM11*, *FUM16*, *FUM17,* and *FUM18*) show some degree of local synteny conservation among the four species that do not contain *FUM* clusters. In other words, each of these genes is in a conserved location among some of the four species, and these locations are not close to one another.

The phylogenetic trees obtained from individual *FUM* genes are shown in Figure 2 (for *FUM*6 and *FUM*15) and supplemental 1 available on line at doi:10.4061/2011/423821 (for the other genes). The trees present a diversity of topologies, such that no one sentence story can explain the origin of the *FUM* cluster. One cannot expect all the *FUM* gene trees to have identical topologies—especially given the possibility that different genes have been subject to very different evolutionary constraints—but even still the diversity of topologies is surprising. To interpret these trees, we consider four possible scenarios for the origin of the cluster (Figure 3), and what tree topologies they would predict. To evaluate these scenarios, we concentrate on the *FUM* genes that are present in both *F. verticillioides* and *A. niger* (*FUM1, FUM6, FUM7, FUM8, FUM9, FUM10, FUM13, FUM14, FUM15,* and *FUM19*). Below, we discuss these four scenarios.

Scenario 1 (*vertical inheritance of an ancestral cluster*). This scenario is illustrated schematically in Figure 3(a). According to it, a cluster existed in the common ancestor of Sordariomycetes (i.e., *F. verticillioides*, *F. graminearum*, *N. crassa,* and *M. grisea*). This cluster became duplicated in this ancestor, and then one copy disintegrated, dispersing its genes around the genome. *F. verticillioides* retained both the cluster and the scattered genes, whereas the other Sordariomycete species retained only the scattered genes.

Support for this scenario comes from some trees (Fum6, Fum10, Fum15, and Fum19) which show that the *FUM* genes have duplicates in *F. verticillioides*, but are single copy in non-producing fumonisin species. A duplication in the common ancestor of Sordariomycetes is suggested by the trees for some genes (Fum6, Fum13, and Fum15), whereas an older duplication in the common ancestor of Sordariomycetes plus Eurotiomycetes (not as shown in Figure 3(a)) is suggested by the trees for other genes (Fum7, Fum8, Fum10, and Fum19).

Scenario 2 (*ancient duplications of scattered genes, followed by recent assembly of a cluster in F. verticillioides*). This scenario is illustrated in Figure 3(b). Scenarios 1 and 2 both require that through evolution numerous independent events of loss have occurred in very distantly related species (for illustration purposes all the losses in Figures 3(a) and 3(b) are placed on the *F. graminearum* branch and in the common ancestor of *N. crassa* and *M. graminearum*, but other combinations or losses on other branches are also possible).

One problem with Scenarios 1 and 2 is that, according to the phylogenies of Fum7, Fum10, Fum15, and Fum19,

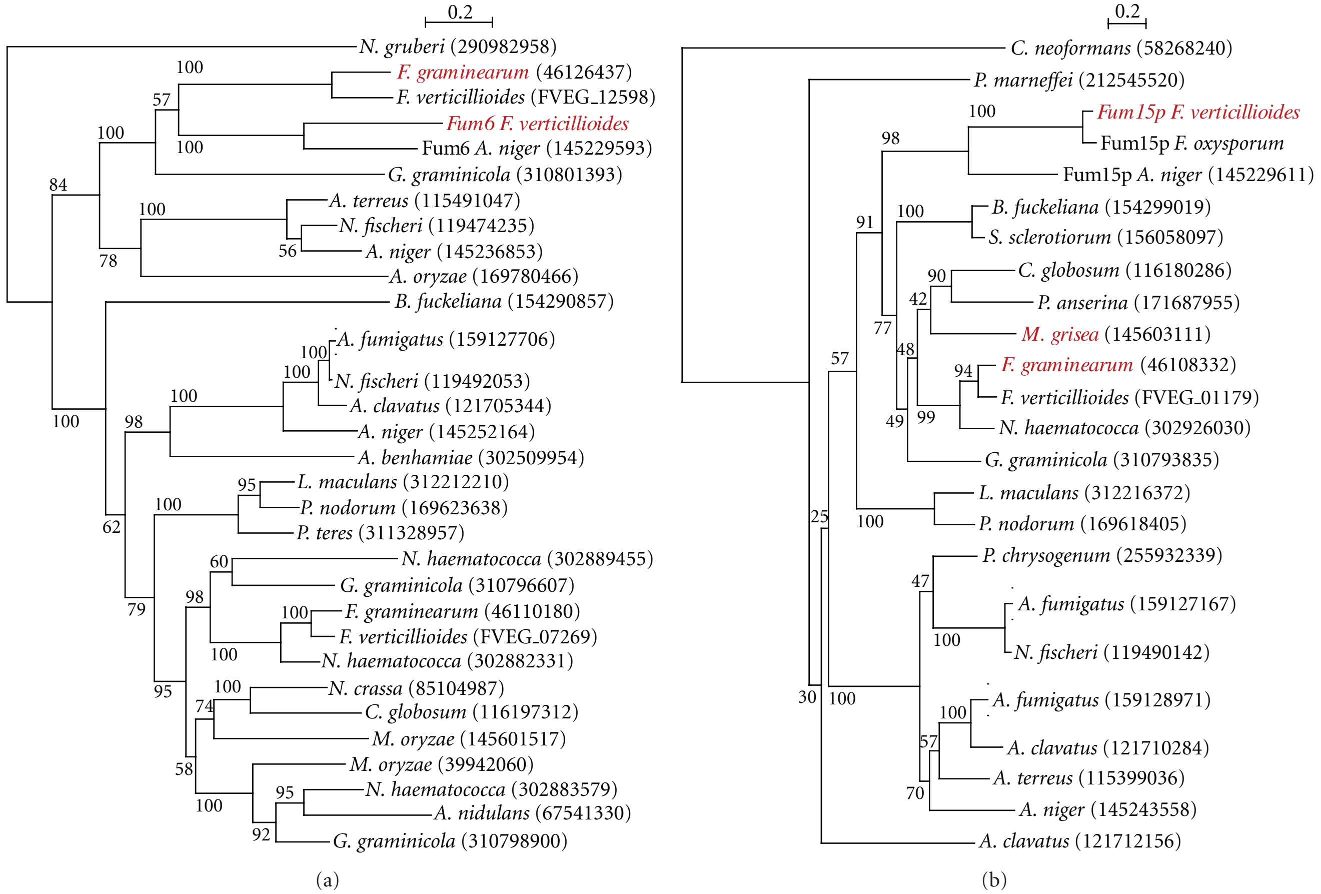


Figure 2: Maximum likelihood trees for FUM6, 15, and their homologs. (a) FUM6; (b) FUM15. In each tree, genes that appear in Figure 1 are named in red. The species name and the NCBI ID are provided on each branch. Bootstrap percentages are shown for all nodes. Trees were constructed from amino acid sequences as described in Section 2 using PHYML after alignment with ClustalW.

the duplicates retained in all the fumonisin-nonproducing Sordariomycetes are coincidently always the same copy (as illustrated by the parallel losses of multiple green genes, but not pink ones, in Figures 3(a) and 3(b)). This can be visualized in the trees by the fact that the homologs of the *FUM* genes in fumonisin-non-producing species are orthologs of each other, showing a more or less typical species phylogeny. If the second copy in *F. graminearum*, *M. grisea*, or *N. crassa* (the one represented in green dots in Figures 3(a) and 3(b) had been retained, this copy would be closer to the *FUM* gene than to genes in the fumonisin-non-producing species, which is not what we observe. Together these observations make both Scenarios 1 and 2 very unlikely.

Scenario 3 (*FUM gene duplication and cluster assembly specifically on the branch leading to F. verticillioides and the GFSC*). The specificity of the *FUM* cluster to the *Gibberella fujikuroi* species complex (GFSC, which includes *F. verticillioides*, *F. oxysporum*, and *F. proliferatum*) points towards a complex-specific cluster. This scenario is shown in Figure 3(c). It proposes that the FUM cluster was built in an ancestor of *F. verticillioides* after its speciation from *F. graminearum*. Most of the gene trees do not support this model, because with the exception of two genes (Fum10 and Fum19) no homolog of a *FUM* gene has remained in *F. verticillioides*. Although the trees for Fum6 and Fum15 show homologs in *F. verticillioides*, the duplications in these cases greatly precede the origin of the GFSC clade.

The numbers of steps required for the different scenarios in Figures 3(a), 3(b), and 3(c) make it more parsimonious to argue that the *FUM* cluster became assembled in an ancestor of *F. verticillioides* (Figure 3(c)) than to argue that either the cluster or the individual genes underwent early duplication and then got lost multiple times (Figures 3(a) and 3(b)). Additionally, as explained above, Figures 3(a) and 3(b) would also imply that the same copy of the duplicates in the fumonisin-non-producing species were independently retained (a minimum of two independent retentions of the same copy are required: one on the *F. graminearum* lineage, and one in the common ancestor of *N. crassa* and *M. grisea*, pink genes in Figures 3(a) and 3(b)).

The hypothesis of assembly requires that each gene transposed once, from an ancestral location, to its current location in *F. verticillioides*. The genes must have been sequentially relocated, with selection for each step.

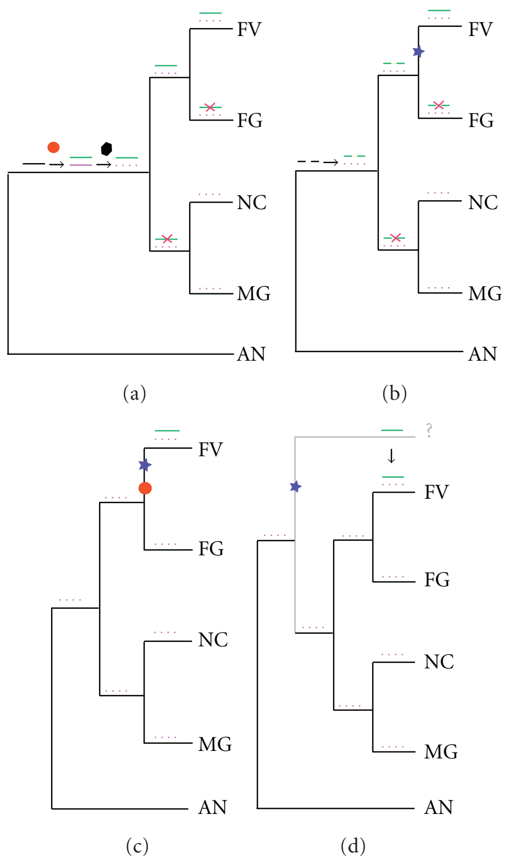


FIGURE 3: The four most likely scenarios giving rise to the current fumonisin cluster in *F. verticillioides*. The species represented on all four trees are: *F. verticillioides* (FV), *F. graminearum* (FG), *N. crassa* (NCU), *M. grisea* (MG), and *A. nidulans* (AN). A red circle represents a duplication event. A blue star represents an assembly and clustering event, while a disassembly is shown using a black square. Red crosses indicate a loss event. When the FUM genes, or their ancestral genes, are clustered they are represented by a short line. The line is dashed if the genes are not clustered. Two colors are assigned to the duplicated genes and green genes are ancestors of the FUM genes (or the current FUM genes), while pink genes are the paralogs of FUM genes (ones found in all other Sordariomycetes). (a) represents the vertical transfer where the ancestor processed a version of the FUM cluster. (b) The ancestor in this scenario contained the ancestral genes of the FUM cluster (scattered). (c) represents a recent event of duplication and assembly of the FUM cluster in an ancestor of F. verticillioides. Finally, (d) represents the horizontal gene transfer scenario.

Scenario 4 (*origin of the F. verticillioides fum cluster by horizontal transfer from a distantly related fungus*). This scenario is shown in Figure 3(d). Under this scenario the donor could be a Sordariomycete (as suggested by the trees for Fum6, Fum13, and Fum15), or a more distant species that is an outgroup to both Sordariomycetes and Eurotiomycetes (as suggested by Fum7, Fum9, Fum10, and Fum19). Although we cannot identify a specific donor, we cannot rule out this possibility. Indeed this scenario reduces the number of events leading to the current trees. Under this scenario, we would not have multiple losses, nor the unexpected retention of the same copy in many of the trees (represented in pink in Figure 3(b); Fum7, Fum10, Fum15, and Fum19). This scenario may explain the unexpected phylogenetic positioning of some *FUM* gene outside the expected class of species. For example, in the case of Fum7, Fum 9, Fum 10, Fum13, Fum15, and Fum19 a horizontal gene transfer of the FUM genes into *F. verticillioides* would explain such topologies.

Figure 3(d) illustrates how the horizontal transfer scenario, like the recent assembly hypothesis (Figure 3(c)), reduces the number of independent events necessary to explain the *FUM* cluster. However, the horizontal transfer scenario does not posit any mechanism for the assembly of the cluster; it just shifts the question to how the cluster became assembled in the donor species.

3.2. *The Fumonisin Cluster in A. niger Results from Horizontal Gene Transfer.*

A. niger has been shown to produce fumonisin [13] and contains clustered homologs of many of the *F. verticillioides* FUM genes [21] (Figure 2). Our phylogenetic analysis illuminates the origin of this cluster in *A. niger* and how it relates to the cluster in *F. verticillioides* (phylogenetic trees in Figure 2 and Supplemental Figure 1).

The first trend evident from this phylogenetic analysis is that genes from the *FUM* cluster in *F. verticillioides*, *F. oxysporum*, and *A. niger* define clades supported by high bootstrap values (>90%, Figure 2), to the exclusion of homologous genes from Sordariomycetes and Eurotiomycetes.

Because we extended our analysis to many species, we were faced with the problem of low bootstrap values for many of the *FUM* genes trees. The two trees shown in Figure 2(b) (*FUM6* and *FUM15*) are the ones with the high bootstrap support for relevant branches, and a reasonably correct species phylogeny. Our analysis shows that both these genes in *A. niger* clearly group with genes from the Sordariomycetes, rather than with genes in the more closely related (Eurotiomycetes) species including other *A. niger* genes. Bootstrap values for grouping the *A. niger FUM* genes with the Sordariomycete homologs are 98–100% (Figure 2).

The disagreement of this result with the expected *A. niger* species relationships, are suggestive of horizontal gene transfer between *A. niger* and an ancestor existing prior to the divergence of *F. verticillioides* and *F. oxysporum*. More importantly, it is more likely that the transfer occurred from Sordariomycetes to *A. niger* (or an ancestor of this species), rather than the opposite. Indeed, the opposite would result in the *FUM* genes (from *F. verticillioides*, *F. oxysporum*, and *A. niger*) clustering in the Eurotiomycetes subphylum as opposed to the Sordariomycetes as seen in Figure 2. For the *FUM6* and *FUM15* trees, we used the likelihood ratio test (LRT) to test whether the topologies shown (Figure 2) have significantly higher likelihoods than alternative trees where the *A. niger* was placed in the Eurotiomycetes and constrained to form a monophyletic group. In both cases the topology shown in Figure 2 is significantly more likely than the tree expected if genes were inherited vertically ($P < .001$ for each).

Because the clusters in *A. niger* and in *F. verticillioides* share only 11 of the 17 known *FUM* genes (including *FUM21*), these two types of cluster have probably had a long history of independent evolution, although they certainly share a common ancestor. We conclude that the cluster in *A. niger* originated by horizontal transfer from an ancestor of *F. verticillioides* and *F. oxysporum*.

4. Discussion

In the literature three scenarios for the creation of a gene cluster have been described: horizontal of an existing cluster from one genome to another [30, 31]; the duplication of an ancestral cluster [31]; the *de novo* creation of a cluster from initially scattered genes that become relocated into one locus [4]. We find that the *FUM* genes are apparent duplicates of conserved genes in Sordariomycetes (Figures 1 and 2). We think that two of the scenarios we discussed could plausibly account for the observed data. First, the *FUM* cluster could be the result of horizontal cluster transfer into an ancestor of *F. verticillioides* (Scenario 4). This scenario is similar to our observation of the *ACE1* cluster in *A. clavatus* [31], and the more recent observation of the horizontal gene transfer of the sterigmatocystin cluster [32]. Secondly, the *FUM* cluster may have been assembled after recent gene duplication in an ancestor of *F. verticillioides* (Scenario 3). The latter scenario resembles our previous observations on the *DAL* gene cluster of *S. cerevisiae* [4], though it should be noted that the *DAL* genes code for a catabolic pathway (degradation of allantoin, a secondary nitrogen source), whereas the *FUM* genes are part of an anabolic pathway (secondary metabolite biosynthesis).

On the other hand we propose that the fumonisin cluster in *A. niger* was acquired via horizontal gene transfer. It has been shown in recent years that horizontal gene transfer between filamentous fungi is more common than was originally thought. Many independent genes can transfer between distantly related species such as that observed between and ancestor of *A. oryzae* and Sordariomycetes [33]; also an entire secondary metabolite cluster has been shown to have horizontally transferred between a relative of *M. grisea* into an ancestor of *A. clavatus* [31]. Here again this finding adds to the repertoire of horizontally transferred genes between fungal species and shows that this exchange mechanism is not so uncommon after all. Moreover it shows how an entire cluster can transfer between distantly related species and remain functional in the new species. In addition, the differences between the *A. niger* and *F. verticillioides* Fum clusters highlights how a cluster can diverge by adding, removing, or reshuffling the genes.

Our lack of knowledge about what benefit the metabolite confers on the organism hampers our understanding of the selective purpose of this clustering. However, we can be almost certain that the reason behind the clustering is not simply to synthesize the metabolite, which is possible with scattered genes. It is more likely that the selective force involves selection for a tight coregulation of gene expression, perhaps mediated by a LaeA-type universal regulator.

Competition, either between one fungal species and another, or between a fungus and a host species, is likely to result in strong selection on the secondary metabolite repertoire of filamentous fungal species. This arms race between organisms pressurizes the organism to create new chemical weapons, which are the products of new secondary metabolite gene clusters. It is relatively easy to envisage that neofunctionalization after gene duplication, or partial cluster duplication as appears to have happened in the origins of the Ace1 cluster [31], could result in the production of a new secondary metabolite and so could be selectively advantageous. It is harder to understand why relocating genes, as has happened in the FUM cluster, can be evolutionarily advantageous. One possibility is that the mere act of relocating a gene can have the consequence of changing the end product of a pathway, because the expression of all the genes in a cluster is coordinated. For example, imagine that we have two secondary metabolite biosynthesis pathways, 1 and 2. If a cytochrome P450 oxidoreductase gene that originally functioned in pathway 1 is suddenly relocated so that it becomes coexpressed with the genes in pathway 2 (and no longer co-expressed with pathway 1), it is possible that its product could begin to act on one of the intermediate molecules in pathway 2. The result would be that the products of pathways 1 and 2 are both changed. Alternative possibilities include that there is selection for tighter regulation (e.g., if an intermediate molecule in the pathway is toxic), or that there is epistatic selection for tight linkage between interacting alleles [34].

Acknowledgment

The authors would like to thank the reviewers for their valuable comments and suggestions.

References

[1] L. D. Hurst, C. Pál, and M. J. Lercher, “The evolutionary dynamics of eukaryotic gene order,” *Nature Reviews Genetics*, vol. 5, no. 4, pp. 299–310, 2004.

[2] L. D. Hurst, E. J. B. Williams, and C. Pál, “Natural selection promotes the conservation of linkage of co-expressed genes,” *Trends in Genetics*, vol. 18, no. 12, pp. 604–606, 2002.

[3] G. A. C. Singer, A. T. Lloyd, L. B. Huminiecki, and K. H. Wolfe, “Clusters of co-expressed genes in mammalian genomes are conserved by natural selection,” *Molecular Biology and Evolution*, vol. 22, no. 3, pp. 767–775, 2005.

[4] S. Wong and K. H. Wolfe, “Birth of a metabolic gene cluster in yeast by adaptive gene relocation,” *Nature Genetics*, vol. 37, no. 7, pp. 777–782, 2005.

[5] B. Field and A. E. Osbourn, “Metabolic diversification—independent assembly of operon-like gene clusters in different plants,” *Science*, vol. 320, no. 5875, pp. 543–547, 2008.

[6] X. Qi, S. Bakht, M. Leggett, C. Maxwell, R. Melton, and A. Osbourn, “A gene cluster for secondary metabolism in oat: implications for the evolution of metabolic diversity in plants,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 21, pp. 8233–8238, 2004.

[7] N. H. Giles, M. E. Case, and J. Baum, “Gene organization and regulation in the qa (quinic acid) gene cluster of Neurospora

crassa," *Microbiological Reviews*, vol. 49, no. 3, pp. 338–358, 1985.

[8] E. P. Hull, P. M. Green, H. N. Arst, and C. Scazzocchio, "Cloning and physical characterization of the L-proline catabolism gene cluster of Aspergillus nidulans," *Molecular Microbiology*, vol. 3, no. 4, pp. 553–559, 1989.

[9] N. P. Keller and T. M. Hohn, "Metabolic pathway gene clusters in filamentous fungi," *Fungal Genetics and Biology*, vol. 21, no. 1, pp. 17–29, 1997.

[10] J. W. Cary, P.-K. Chang, and D. Bhatnagar, "Clustered metabolic pathway genes in filamentous fungi," in *Applied Mycology and Biotechnology*, G. G. Khachatourians and D. K. Arora, Eds., pp. 165–198, Elsevier, Amsterdam, The Netherlands, 2001.

[11] D. W. Brown, J. H. Yu, H. S. Kelkar et al., "Twenty-five coregulated transcripts define a sterigmatocystin gene cluster in Aspergillus nidulans," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 93, no. 4, pp. 1418–1422, 1996.

[12] R. H. Proctor, D. W. Brown, R. D. Plattner, and A. E. Desjardins, "Co-expression of 15 contiguous genes delineates a fumonisin biosynthetic gene cluster in Gibberella moniliformis," *Fungal Genetics and Biology*, vol. 38, no. 2, pp. 237–249, 2003.

[13] D. W. Brown, R. A. E. Butchko, M. Busman, and R. H. Proctor, "The *Fusarium verticillioides* FUM gene cluster encodes a Zn(II)2Cys6 protein that affects FUM gene expression and fumonisin production," *Eukaryotic Cell*, vol. 6, no. 7, pp. 1210–1218, 2007.

[14] D. W. Brown, R. B. Dyer, S. P. McCormick, D. F. Kendra, and R. D. Plattner, "Functional demarcation of the Fusarium core trichothecene gene cluster," *Fungal Genetics and Biology*, vol. 41, no. 4, pp. 454–462, 2004.

[15] N. P. Keller, G. Turner, and J. W. Bennett, "Fungal secondary metabolism—from biochemistry to genomics," *Nature Reviews Microbiology*, vol. 3, no. 12, pp. 937–947, 2005.

[16] U. L. Rosewich and H. C. Kistler, "Role of horizontal gene transfer in the evolution of fungi," *Annual Review of Phytopathology*, vol. 38, pp. 325–363, 2000.

[17] N. Khaldi, F. T. Seifuddin, G. Turner et al., "SMURF: genomic mapping of fungal secondary metabolite clusters," *Fungal Genetics and Biology*, vol. 47, no. 9, pp. 736–741, 2010.

[18] C. Waalwijk, T. Van Der Lee, I. De Vries, T. Hesselink, J. Arts, and G. H. J. Kema, "Synteny in toxigenic Fusarium species: the fumonisin gene cluster and the mating type region as examples," *European Journal of Plant Pathology*, vol. 110, no. 5-6, pp. 533–544, 2004.

[19] R. H. Proctor, R. D. Plattner, D. W. Brown, J. A. Seo, and Y. W. Lee, "Discontinuous distribution of fumonisin biosynthetic genes in the Gibberella fujikuroi species complex," *Mycological Research*, vol. 108, no. 7, pp. 815–822, 2004.

[20] J. C. Frisvad, J. Smedsgaard, R. A. Samson, T. O. Larsen, and U. Thrane, "Fumonisin B production by Aspergillus niger," *Journal of Agricultural and Food Chemistry*, vol. 55, no. 23, pp. 9727–9732, 2007.

[21] S. E. Baker, "Aspergillus niger genomics: past, present and into the future," *Medical Mycology*, vol. 44, no. 1, pp. 17–21, 2006.

[22] J. E. Galagan, S. E. Calvo, C. Cuomo et al., "Sequencing of Aspergillus nidulans and comparative analysis with A. fumigatus and A. oryzae," *Nature*, vol. 438, no. 7071, pp. 1105–1115, 2005.

[23] C. A. Cuomo, U. Güldener, J.-R. Xu et al., "The Fusarium graminearum genome reveals a link between localized polymorphism and pathogen specialization," *Science*, vol. 317, no. 5843, pp. 1400–1402, 2007.

[24] R. A. Dean, N. J. Talbot, D. J. Ebbole et al., "The genome sequence of the rice blast fungus Magnaporthe grisea," *Nature*, vol. 434, no. 7036, pp. 980–986, 2005.

[25] J. E. Galagan, S. E. Calvo, K. A. Borkovich et al., "The genome sequence of the filamentous fungus Neurospora crassa," *Nature*, vol. 422, no. 6934, pp. 859–868, 2003.

[26] D. W. Brown, F. Cheung, R. H. Proctor et al., "Comparative analysis of 87,000 expressed sequence tags from the fumonisin-producing fungus Fusarium verticillioides," *Fungal Genetics and Biology*, vol. 42, no. 10, pp. 848–861, 2005.

[27] J. D. Thompson, D. G. Higgins, and T. J. Gibson, "CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice," *Nucleic Acids Research*, vol. 22, no. 22, pp. 4673–4680, 1994.

[28] J. Castresana, "Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis," *Molecular Biology and Evolution*, vol. 17, no. 4, pp. 540–552, 2000.

[29] S. Guindon and O. Gascuel, "A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood," *Systematic Biology*, vol. 52, no. 5, pp. 696–704, 2003.

[30] N. J. Patron, R. F. Waller, A. J. Cozijnsen et al., "Origin and distribution of epipolythiodioxopiperazine (ETP) gene clusters in filamentous ascomycetes," *BMC Evolutionary Biology*, vol. 7, article 174, 2007.

[31] N. Khaldi, J. Collemare, M. H. Lebrun, and K. H. Wolfe, "Evidence for horizontal transfer of a secondary metabolite gene cluster between fungi," *Genome Biology*, vol. 9, no. 1, article R18, 2008.

[32] J. C. Slot and A. Rokas, "Horizontal transfer of a large and highly toxic secondary metabolic gene cluster between fungi," *Current Biology*, vol. 21, no. 2, pp. 134–139, 2011.

[33] N. Khaldi and K. H. Wolfe, "Elusive origins of the extra genes in Aspergillus oryzae," *PLoS ONE*, vol. 3, no. 8, article e3036, 2008.

[34] M. Nei, "Modification of linkage intensity by natural selection," *Genetics*, vol. 57, no. 3, pp. 625–641, 1967.

SAGE-Hindawi Access to Research
International Journal of Evolutionary Biology
Volume 2011, Article ID 143498, 11 pages
doi:10.4061/2011/143498

Research Article

Computational Analysis Suggests That Lyssavirus Glycoprotein Gene Plays a Minor Role in Viral Adaptation

Kevin Tang[1] and Xianfu Wu[2]

[1] *BCFB, DSR, Centers for Disease Control and Prevention, Atlanta, GA 30333, USA*
[2] *Rabies, PRB, Centers for Disease Control and Prevention, Atlanta, GA 30333, USA*

Correspondence should be addressed to Xianfu Wu, xwu@cdc.gov

Received 15 October 2010; Revised 15 December 2010; Accepted 3 January 2011

Academic Editor: Hiromi Nishida

The Lyssavirus glycoprotein (G) is a membrane protein responsible for virus entry and protective immune responses. To explore possible roles of the glycoprotein in host shift or adaptation of *Lyssavirus*, we retrieved 53 full-length glycoprotein gene sequences from NCBI GenBank. The sequences were from different host isolates over a period of 70 years in 21 countries. Computational analyses detected 1 recombinant (AY987478, a dog isolate of CHAND03, genotype 1 in India) with incongruent phylogenetic support. No recombination was detected when AY98748 was excluded in the analyses. We applied different selection models to identify selection pressure on the glycoprotein gene. One codon at amino acid residual 483 was found to be under weak positive selection with marginal probability of 95% by using the maximum likelihood method. We found no significant evidence of positive selection on any site of the glycoprotein gene when the putative recombinant AY987478 was excluded. The computational analyses suggest that the G gene has been under purifying selection and that the evolution of the G gene may not play a significant role in Lyssavirus adaptation.

1. Introduction

Positive selection and recombination are important mechanisms in microbial pathogen adaption to new hosts, resistance to antibiotics, and evasion of immune responses [1]. RNA viruses have high mutation rates due to lack of both proofreading and postreplicative repair activities associated with RNA replicases and reverse transcriptases [2], which benefits RNA viruses in adapting to the changing environment. Recombination is a general phenomenon in evolution and plays a significant role in viral fitness [3, 4]. Rabies virus is a single-stranded negative RNA virus belonging to the order *Mononegavirales*, family *Rhabdoviridae*, genus *Lyssavirus*, which causes rabies in all warm-blooded mammals. Host shift and spillover events are frequently reported in rabies [5–9]. The nucleotide substitution rate of lyssaviruses is estimated to be around 10^{-4} per site per year [7]. The RNA-dependent RNA-polymerase (RdRp or L) together with phosphoprotein (P), functions as the transcriptase and replicase complex. The glycoprotein (G) is the only outer membrane protein responsible for virus entry and inducing protective immune responses [10, 11]. The role of the G gene in rabies spillover, host shift, and adaptation has not been analyzed thoroughly. The information could help understand viral pathogenesis and develop a vaccine for a broad spectrum of lyssavirus infections.

Here, we used newly developed computational algorithms as well as traditional methods to investigate potential recombination events and selection pressures in the G gene of Lyssaviruses. The dataset for the study was comprised of 53 full-length glycoprotein gene sequences isolated from different hosts in 21 countries over a period of 70 years. We hypothesized that if different hosts with rabies infections over decades did not lead to positive selection or recombination events in the G gene, the gene does not play a significant role in lyssavirus adaptation.

2. Methods

2.1. Dataset. We choose a dataset that covers lyssavirus isolates spatially and geographically over a long period of time in various animal hosts. Fifty-three full-length G sequences from 21 countries isolated over a period of 70 years were retrieved from NCBI GenBank. The sequences were aligned using fast statistical alignment (FSA, [12]). Briefly, FSA is a probabilistic multiple-sequence alignment algorithm, which uses a "distance-based" approach to aligning homologous protein, RNA, or DNA sequences. It produces superior alignments of homologous sequences that are subject to very different evolutionary constraints. The nucleotide (nt) sequence alignment of the lyssavirus G genes was corrected manually by visual inspection using the amino acid sequence alignment. Gaps were removed if they existed in majority of the sequences.

2.2. Phylogenetic Analyses. A phylogenetic tree was reconstructed by using the neighbor joining algorithm in the MEGA 4 package [13]. The maximum composite likelihood model was used as well as the pairwise deletion option for gaps. The statistical significance of the phylogeny was measured by bootstrap with 1,000 replicates.

2.3. Recombination Detection. We first applied PHI [14], NSS [15], and Max χ^2 [16] tests (implemented in PhiPack [14]) with 1,000 permutations to detect recombination. Sequences involved in the recombination and breakpoints were determined by using 3SEQ [17] and GARD implemented in the Datamonkey web interface [18, 19]. The recombination was further verified by bootscanning and phylogenetic incongruence analysis. Bootscanning was performed using SimPlot software version 3.5.1 [20]. The parameters for bootscanning were window size, 200 bp; step, 10 bp; GapStrip, on; bootstrap replicate, 1000; distance model, Kimura (2-parameter); tree algorithm, neighbor-joining.

2.4. Selection Analyses. To test positive selection on sites of the G gene in Lyssaviruses, the Codeml program in PAML software package version 4.4 was employed [21]. Codeml implements the maximum likelihood method to test if positive selection has taken place at sites within a gene. This method uses different codon substitution models to estimate the number of nonsynonymous (dN) and synonymous substitutions (dS) per site among codons, since different amino acids in a protein could be under different selective pressures, thus creating a different ω (dN/dS) ratio. The models in our dataset analyses were M0 (one-ratio), M1 (nearly neutral), M2 (positive selection), M7 (β distribution), and M8 ($\beta + \omega > 1$) [22]. The M0 model estimates overall ω for the data. The M1 model estimates codon site proportion p_0 with $\omega_0 < 1$ and proportion p_1 ($p_1 = 1 - p_0$) with $\omega_1 = 1$. The M2 model allows an additional class of positively selected sites with proportion p_2 ($p_2 = 1 - p_1 - p_0$) with ω_2 estimated from the data. The M7 model specifies that ω follows a beta distribution and the value of ω is allowed to change between 0 and 1. Parameters p and q of the beta distribution are estimated from the data in the M7 model. In the M8 model, a proportion of sites p_0 has a ω in the beta distribution and the proportion p_1 sites are assumed to be positively selected. Two sets of comparisons (M2 versus M1, M8 versus M7) were made to test the hypothesis of selection. Within the comparison, the likelihood ratio test statistic used to determine the level of significance was calculated as twice the difference of the likelihood scores ($2\Delta l$) estimated by each model. The significance was determined under χ^2 distribution. The degrees of freedom for the M1 versus M2 and M7 versus M8 tests are 2 [22]. If M8 or M2 is significantly favored and it contains codons with $\omega > 1$, positive selection is significantly evident. Posterior probabilities of the inferred positively selected sites were estimated by the Bayes empirical Bayes (BEB) approach [23].

We also applied single-likelihood ancestor counting (SLAC), fixed-effects likelihood (FEL), and random-effects likelihood (REL) [18] to indentify selection pressure on individual codons of the G gene in lyssaviruses.

3. Results

3.1. Recombination Analyses. Our dataset covered lyssaviruses isolated over a period of 70 years from 21 countries (Table 1), including the new and old continents. The hosts included bats, cows, dogs, foxes, humans, raccoons, sheep, and skunks.

The PHI and Max χ^2 tests suggested significant evidence of recombination in the G gene. By 1000 permutations, the P-values of PHI and Max χ^2 test were .006 and 0, respectively. However, no significant evidence ($P = .796$) of recombination was detected by using the NSS test.

By using 3SEQ, 6 long recombinant sequences (>100 bp) were detected: AF233275, AY237121, AY987478, DQ074978, DQ849071, and L04523 (Table 2). Two breakpoints were identified in all recombinants. The first breakpoint was at nucleotide position between 400 and 800. The second breakpoint was around nucleotide position of 1080. However, the two breakpoints for DQ074978 and L04523 were at the very beginning and around nucleotide position of 109, respectively.

The analysis by using GARD also suggested evidence of recombination with significant topological incongruence at the 2 breakpoints (Table 3). The first breakpoint was at nucleotide position of 441 and the second was at nucleotide position of 1089. The significance value for the 2 breakpoints was 0.01. The left hand side (LHS) and the right hand side (RHS) P-values for the 2 breakpoints were .0004.

We analyzed the recombination events by using BootScanning as implemented in SimPlot. Sequence AY987478 was used as a query sequence in all four cases (Figures 1(a)–1(d)). The analysis confirmed the recombination event in the G gene of lyssavirus. The high bootstrap values support clustering sequence AY987478 with AF325489 (Figures 1(a) and 1(b)) and with AY237121 (Figures 1(c) and 1(d)) at positions from 1 to around 440 and at positions from around 1130 to the end of the sequences. The bootstrap values are also high for clustering AY987478 with AF23375 (Figures 1(a) and 1(c)) and DQ074978 (Figures 1(b) and 1(d)) at positions from around 540 to 1000. The switches of the high bootstrap values at nucleotide positions from around 440 to 540 and from 1000 to 1130 indicate two possible breakpoints for the recombination.

Table 1: Sequences of glycoprotein gene used in this study.

Accession no.	Country	Host	Year of isolation	Strain/isolate	Genotype	References
AB115921	Indonesia	Dog	2001	SN01-23	GT1	Unpublished
AF233275	India	Sheep		PV11	GT1	Unpublished
AF298141	USA	Bat	1979	USA7-BT	GT1	Badrane et al. [24]
AF298142	Poland	Bat	1985	EBL1POL	GT5	Badrane et al. [24]
AF298143	France	Bat	1989	EBL1FRA	GT5	Badrane et al. [24]
AF298144	Finland	Bat	1986	EBL2FIN	GT6	Badrane et al. [24]
AF298145	Holland	Bat	1986	EBL2HOL	GT6	Badrane et al. [24]
AF298146	S. Africa	Bat	1970	DuvSAF1	GT4	Badrane et al. [24]
AF298147	S. Africa	Bat	1981	DuvSAF2	GT4	Badrane et al. [24]
AF325487	Malaysia	Human	1985	MAL1-HM	GT1	Badrane and Tordo [7]
AF325489	Nepal	Dog	1989	NEP1-DG	GT1	Badrane and Tordo [7]
AF325490	French	Bovine	1985	GUY1-BV	GT1	Badrane and Tordo [7]
AF325491	Brazil	Bovine	1986	BRA1-BV	GT1	Badrane and Tordo [7]
AF325492	Mexico	Bat	1987	MEX2-VP	GT1	Badrane and Tordo [7]
AF325494	USA	Bat	1981	USA8-BT	GT1	Badrane and Tordo [7]
AF325495	USA	Bat	1982	USA9-BT	GT1	Badrane and Tordo [7]
AF401285	Thailand			8743THA	GT1	Unpublished
AF426297	Australia	Bat	1997	ABLSF12NB	GT7	Guyatt et al. [25]
AF426298	Australia	Bat	1997	ABLSF11KW	GT7	Guyatt et al. [25]
AJ871962	China	Vaccine		PM	GT1	Unpublished
AY009098	China	Human	1986	CNX8601	GT1	Tang et al. [26]
AY009099	China	Human	1986	CNX8511	GT1	Tang et al. [26]
AY009100	China	Dog (Vaccine)	1955	CTN	GT1	Tang et al. [26]
AY237121	India	Dog		RVD	GT1	Unpublished
AY257980	Thailand	Human		HM65	GT1	Hemachudha et al. [27]
AY257982	Thailand	Human		HM88	GT1	Hemachudha et al. [27]
AY257983	Thailand	Human		HM208	GT1	Hemachudha et al. [27]
AY987478	India	Dog	1999	CHAND03	GT1	Unpublished
D14873	Japan	Vaccine		RC-HL	GT1	Unpublished
D16330	Japan	Vaccine		RC-HL	GT1	Ito et al. [28]
DQ074978	India	Dog			GT1	Agrawal et al. [29]
DQ076097	S. Korea	Bovine		SKRBV0404HC	GT1	Hyun et al. [30]
DQ076099	S. Korea	Dog		SKRRD9903YG	GT1	Hyun et al. [30]
DQ767897	China	Vaccine		CTN-35	GT1	Unpublished
DQ849071	China	Dog	1994	GX4	GT1	Meng et al. [31]
DQ849072	China	Dog	1992	CQ92	GT1	Meng et al. [31]
L04522	China	Vaccine (Dog)	1931	3aG	GT1	Bai et al. [32]
L04523	China	Vaccine (dog)	1993	CGX89-1	GT1	Bai et al. [32]
L40426				CVS	GT1	Yelverton et al. [33]
M81058	Algeria	Dog		ALG1-DG	GT1	Benmansour et al. [34]
M81059	Algeria	Human			GT1	Benmansour et al. [34]
M81060	Algeria	Human			GT1	Benmansour et al. [34]
U03765	Canada	Vulpes		8480FX	GT1	Nadin-Davis et al. [35]
U03766	Arctic Circle	Dog	1992	Arctic A1-1090DG	GT1	Nadin-Davis et al. [35]
U03767	Canada	Dog	1993	Hudson Bay-4055DG	GT1	Nadin-Davis et al. [35]
U11736	Canada	Arctic Fox		91RABN1035	GT1	Nadin-Davis et al. [36]
U11755	Canada	Skunk		91RABN1578	GT1	Nadin-Davis et al. [36]
U27214	USA	Raccoon		NY 516	GT1	Nadin-Davis et al. [37]
U27215	USA	Raccoon		NY 771	GT1	Nadin-Davis et al. [37]
U27216	USA	Raccoon		FLA 125	GT1	Nadin-Davis et al. [37]
U27217	USA	Raccoon		PA R89	GT1	Nadin-Davis et al. [37]
U52946	USA	Bat	1994	SHBRV	GT1	Morimoto et al. [38]
X69122	India	Vaccine		Flury	GT1	Unpublished

Table 2: Recombination detection in glycoprotein gene of lyssavirus by using 3SEQ.

P	Q	C	*P*-value	Dunn Sidak	Breakpoints	
M81058	AY987478	AF233275	0	2.08*E* − 08	432–440, 1080–1089	456–496, 1080–1089
M81060	AY987478	AF233275	1*E* − 12	1.31*E* − 07	432–440, 1080–1089	456–496, 1080–1089
AY987478	M81059	AY237121	0	2.81*E* − 11	441–455, 1077–1079	
AY987478	M81058	AY237121	0	2.13*E* − 13	441–455, 1077–1079	
AY987478	M81060	AY237121	0	1.13*E* − 13	441–455, 1077–1079	
AY987478	AF233275	AY237121	1.3*E* − 10	1.88*E* − 05	432–455, 1068–1089	465–518, 1068–1089
AY987478	L04522	AY237121	1.1*E* − 08	1.48*E* − 03	627–638, 1077–1089	663–666, 1077–1089
AY987478	AF325489	AY237121	0	2.71*E* − 15	700-701, 1077–1097	
AY987478	U11755	AY237121	3.2*E* − 10	4.42*E* − 05	717–719, 1077–1082	729–734, 1077–1082
AY987478	U11736.2	AY237121	3.3*E* − 09	4.61*E* − 04	717–719, 1077–1082	729–734, 1077–1082
AY987478	DQ849071	AY237121	6.1*E* − 11	8.61*E* − 06	736-737, 1077–1079	
AY987478	DQ076097	AY237121	1.2*E* − 10	1.69*E* − 05	630–638, 1077–1089	699–701, 1077–1089
AY987478	DQ076099	AY237121	9*E* − 12	1.31*E* − 06	700-701, 1077–1089	714–719, 1077–1089
AY987478	L04523	AY237121	2.2*E* − 09	3.04*E* − 04	736-737, 1077–1079	
AY987478	X69122	AY237121	4*E* − 12	6.00*E* − 07	666–669, 1032–1049	666–669, 1077–1089
AY987478	AY009098	AY237121	4*E* − 12	4.99*E* − 07	693–701, 1077–1079	705–711, 1077–1079
AY987478	AY009099	AY237121	4*E* − 12	4.99*E* − 07	693–701, 1077–1079	705–711, 1077–1079
AY987478	DQ849072	AY237121	2.1*E* − 11	3.02*E* − 06	693–701, 1077–1079	705–711, 1077–1079
AY987478	AJ871962	AY237121	1*E* − 12	7.29*E* − 08	750–794, 1077–1089	
AY987478	AF325487	AY237121	0	1.36*E* − 08	780–794, 1077–1079	
AY987478	L40426	AY237121	4.8*E* − 11	6.71*E* − 06	750–794, 1077–1089	
AY987478	AF401285	AY237121	0	9.99*E* − 10	780–795, 1077–1079	
AY987478	AY257983	AY237121	2.3*E* − 11	3.27*E* − 06	780–795, 1077–1079	
AY987478	AY257980	AY237121	0	5.70*E* − 09	750–761, 1077–1079	780–795, 1077–1079
AY987478	AY257982	AY237121	5.9*E* − 11	8.33*E* − 06	780–795, 1032–1043	780–795, 1077–1079
AY987478	DQ767897	AY237121	1*E* − 07	1.46*E* − 02	759–767, 972–974	
AY987478	U52946	AY237121	5.3*E* − 08	7.43*E* − 03	741–748, 900-901	741–748, 918–938
AY987478	U03766	AY237121	2.4*E* − 07	3.27*E* − 02	717–719, 876–889	717–719, 894–914
AY987478	U03765	AY237121	2.6*E* − 07	3.55*E* − 02	717–719, 876–889	717–719, 894–914
AY237121	AF233275	AY987478	0	1.38*E* − 39	432–452, 1077–1089	
AY237121	DQ074978	AY987478	0	1.16*E* − 38	432–452, 1077–1089	
AY237121	L04522	AY987478	0	7.15*E* − 25	627–647, 1065–1089	
AY237121	DQ076097	AY987478	0	1.47*E* − 08	630–647, 1056–1058	630–647, 1065–1089
AY237121	U03767	AY987478	1*E* − 11	1.42*E* − 06	630–638, 1041–1058	630–638, 1065–1079
AY237121	AJ871962	AY987478	0	6.36*E* − 20	642–647, 1041–1058	642–647, 1065–1089
AY237121	X69122	AY987478	0	3.16*E* − 26	642–647, 1041–1049	654–659, 1041–1049
AY237121	L40426	AY987478	0	1.38*E* − 17	642–647, 1041–1058	642–647, 1065–1089
AY237121	M81058	AY987478	0	9.60*E* − 21	441–452, 1065–1079	618–710, 1065–1079
AY237121	M81060	AY987478	0	6.49*E* − 23	441–452, 1065–1079	618–710, 1065–1079
AY237121	D14873	AY987478	0	6.82*E* − 17	685–701, 1065–1085	705–710, 1065–1085
AY237121	D16330	AY987478	0	5.23*E* − 17	685–701, 1065–1085	705–710, 1065–1085
AY237121	AY257980	AY987478	5.3*E* − 08	7.49*E* − 03	705–710, 1041–1046	
AY237121	DQ849071	AY987478	1.9*E* − 09	2.68*E* − 04	708–710, 1041–1046	
AY237121	L04523	AY987478	1.3*E* − 07	1.85*E* − 02	708–710, 1041–1046	
AY237121	DQ076099	AY987478	0	1.07*E* − 08	634–647, 1056–1058	634–647, 1065–1089
AY237121	U11755	AY987478	1*E* − 12	1.60*E* − 07	630–647, 1056–1058	630–647, 1065–1082
AY237121	U11736.2	AY987478	0	2.29*E* − 08	630–647, 1056–1058	630–647, 1065–1082
AY237121	DQ767897	AY987478	1.7*E* − 09	2.37*E* − 04	708–710, 1017–1022	708–710, 1041–1046
AY237121	AY009098	AY987478	1.8*E* − 08	2.58*E* − 03	705–710, 1041–1046	736-737, 1041–1046
AY237121	AY009099	AY987478	1.8*E* − 08	2.58*E* − 03	705–710, 1041–1046	736-737, 1041–1046

Table 2: Continued.

P	Q	C	P-value	Dunn Sidak	Breakpoints	
AY237121	AF325487	AY987478	$7.5E-10$	$1.06E-04$	705–710, 1041–1046	732–734, 1041–1046
AY237121	U03766	AY987478	$1.2E-09$	$1.75E-04$	630–638, 1041–1058	630–638, 1065–1079
AY237121	U03765	AY987478	$2.3E-10$	$3.26E-05$	630–638, 1041–1058	630–638, 1065–1079
AY237121	M81059	AY987478	0	$1.27E-18$	441–452, 993–998	441–452, 1017–1034
AY237121	AF325490	AY987478	$9.2E-09$	$1.30E-03$	705–710, 993–995	705–710, 1017–1019
AY237121	AF325491	AY987478	$7E-12$	$9.93E-07$	705–710, 993–995	
AY237121	AF325492	AY987478	$9.6E-09$	$1.34E-03$	700-701, 993–995	705–710, 993–995
AY237121	DQ849072	AY987478	$1.3E-07$	$1.83E-02$	705–710, 924–935	705–710, 945–950
AY237121	AY009100	AY987478	$3.4E-07$	$4.62E-02$	708–710, 885–887	708–710, 924–938
AY237121	AF401285	AY987478	$6.4E-09$	$8.95E-04$	736-737, 883–887	
AY237121	AY257983	AY987478	$9.9E-08$	$1.38E-02$	732–734, 883–887	732–734, 1041–1046
M81059	AY987478	DQ074978	0	$9.05E-10$	519–522, 1080–1089	
M81058	AY987478	DQ074978	0	$4.12E-09$	519–522, 1080–1089	
M81060	AY987478	DQ074978	0	$2.80E-08$	519–522, 1080–1089	
AY009100	M81059	DQ849071	$2E-08$	$2.80E-03$	0–3, 108–119	
AY009100	M81058	DQ849071	$3E-08$	$4.20E-03$	0–3, 108–119	
AY009100	M81060	DQ849071	$1.2E-07$	$1.70E-02$	0–3, 108–119	
AY009100	AJ871962	DQ849071	$2.2E-07$	$3.02E-02$	0–3, 108–110	
AY009100	M81059	L04523	$9.6E-09$	$1.34E-03$	0–3, 108–119	
AY009100	M81058	L04523	$1.5E-08$	$2.04E-03$	0–3, 108–119	
AY009100	M81060	L04523	$6.7E-08$	$9.37E-03$	0–3, 108–119	0–3, 139–161
AY009100	AJ871962	L04523	$1.3E-07$	$1.88E-02$	0–3, 108–110	

Note: P and Q are putative parent sequences, and C is the putative child sequence in the recombination.

Table 3: KH tests verify the significance of breakpoints estimated by GARD analysis.

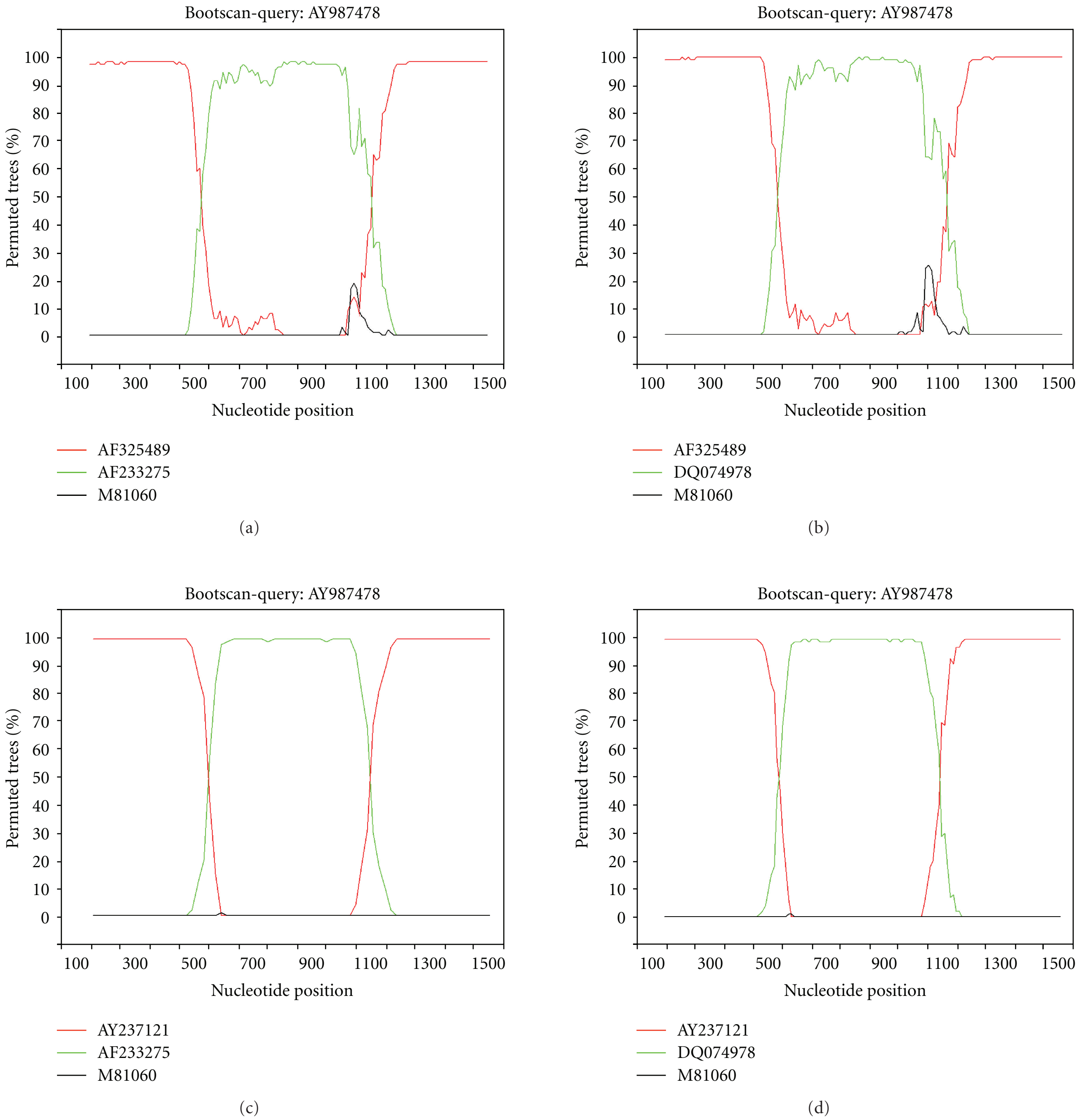
Breakpoint	LHS P-value	RHS P-value	Significance
441	.00040	.00040	0.01
1089	.00040	.00040	0.01

Since recombination with 2 breakpoints was predicted by 3SEQ, GARD, and Bootscanning, we constructed phylogenetic trees by using sequences from the beginning to the first breakpoint and the sequences from the second breakpoint to the end (Figure 2(a)) and a phylogenetic tree with sequences between the two breakpoints (Figure 2(b)). The reconstructed trees presented conflicting topological positions of the putative recombinant AY987478. The putative recombinant was clustered with AY237121 and AF325489 in Figure 2(a), but clustered with DQ074978 and AF233275 in Figure 2(b). All other 5 putative recombinants did not present phylogenetic incongruence. The same result was also verified by GARD (data not shown). When AY987478 was excluded from the dataset, the P-values of Phi, Max χ^2, and NSS were .121, .209, and .791, respectively, suggesting no evidence of recombination. The GARD analysis did not indicate evidence of recombination either.

3.2. Selection Pressure Analyses. The selection pressure analysis with the glycoprotein gene by using PAML is presented in Table 4. The likelihood ratio test statistic ($2\Delta l$) estimated by M2 and M1 was 0. The corresponding P value was .99, which is not significant to reject the nearly null hypothesis of neutral selection in M1. In the comparison between the null neutral site model (M7) and the selection model (M8), the $2\Delta l$ was 18.18 and the corresponding P-value was .0001, indicating that the positive selection model was significantly favored over the null neutral site model. Posterior probabilities of the inferred positively selected sites estimated by the BEB approach were shown in Table 5. Four amino acid sites at 466, 483, 486, and 490 were identified to be under positive selection. But only the site at position 483 had a marginal significance support with posterior probability of 95% and weak positive selection pressure with ω of 1.466. The corresponding posterior probabilities for sites at 466, 486 and, 490 were 68%, 56%, and 82%, respectively.

To test the effect of recombination on positive selection analysis, we excluded the putative recombinant AY987478 from the dataset. Similar results were observed, and the BEB posterior probability supports for amino acid sites under positive selection were nonsignificant (Table 5). When all six putative recombinants were excluded in our analysis, no evidence was found to support positive selection either in M1 or M7 (data not shown). In all cases, the ω in M0 was either 0.07 or 0.08. Overall, 87% of the sites in the G gene had a very low ω value of 0.05 in M2 and M7, indicating strong selective constraints on those sites.

To study the effect of viral passages and possible genetic bottlenecks on the results, we repeated the analysis with a dataset excluding six vaccine sequences and the sequence



(a)

(b)

(c)

(d)

Figure 1: Bootscanning analysis of recombination in glycoprotein gene of lyssavirus by using the SimPlot program with a window size of 200 nucleotides and a step size of 10 nucleotides.

AF233275 (PV11) from cell culture of lyssaviruses under intensive cell culture. We found no significant evidence for positive selection pressure on any site of the G gene.

Analyses using SLAC, REL, and FEL found no evidence of any amino acid in the G gene under positive selection, instead most of the amino acids were found to be under negative selection (Table 6). One site at position 416 was under marginal positive selection by FEL with *P*-value of .0999, narrowly passing the significance level of 0.1. However, this result was not supported by SLAC and REL.

4. Discussion

Lyssaviruses can infect all warm-blooded mammals, and spillover events and host shift have been well documented [5–9]. The molecular mechanism of rabies infection and transmission is still not completely understood, and the phenomenon usually leads to the connection with rabies virus G protein, since G is the only membrane protein responsible for virus entry both in vitro and in vivo. Therefore, it is a reasonable assumption that rabies virus

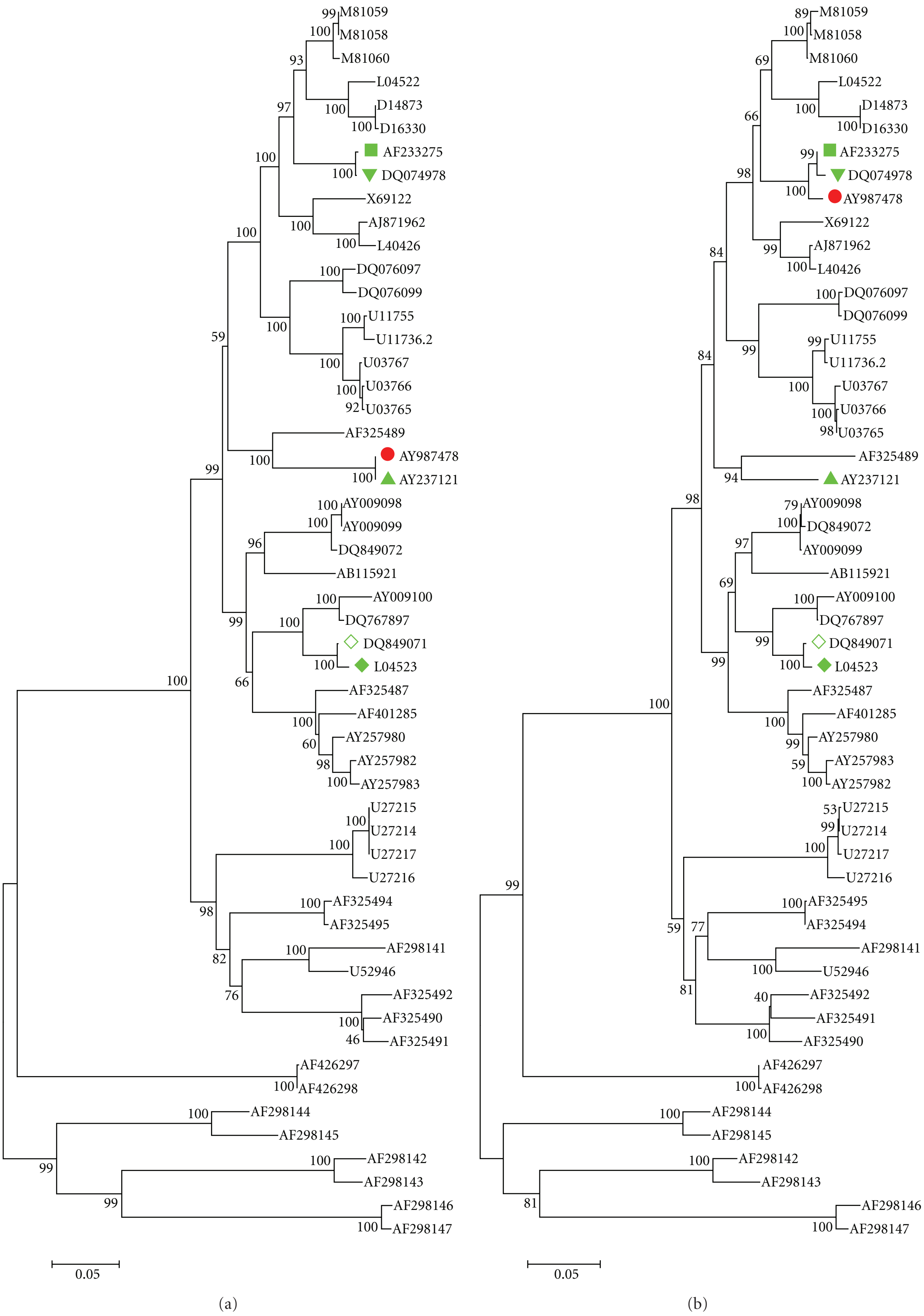


Figure 2: (a) NJ phylogenetic tree of 53 glycoprotein gene sequences with regions concatenated from position of 1 to 441 and position of 1090 to 1572. Bootstrap values of 1000 replicates are shown above the branches. The red marker represents the putative recombinant. (b) NJ phylogenetic tree of 53 glycoprotein gene sequences with region from position of 441 to 1089. Bootstrap values of 1000 replicates are shown above the branches. The red marker represents the putative recombinant.

Table 4: Parameter estimates, *dN/dS* ratio, likelihood score, and test statistics under models of variable ω ratios among sites for the glycoprotein gene in lyssavirus.

	Parameter estimates	*dN/dS*	Likelihood scores (l)	Model comparison ($2\Delta l$, d.f., P)	Positive selection
M0: one ratio	$\omega = 0.08$	0.08	−24586.10		None
M1: Nearly neutral	$\omega_0 = 0.05$, $\omega_1 = 1$, ($p_0 = 0.87$, $p_1 = 0.13$)	0.17	−24010.40		Not allowed
M2: Positive selection	$\omega_0 = 0.05$, $\omega 1 = 1$, $\omega_2 = 1$, ($p_0 = 0.87$, $p_1 = 0.06$, $p_2 = 0.07$)	0.17	−24010.40	M2 versus M1:0, d.f. = 2, $P = .99$	None
M7: β, Neutral	$p = 0.26$, $q = 2.11$	0.10	−23443.16		Not allowed
M8: $\beta + \omega > 1$, Selection	$p_0 = 0.98$, $p = 0.28$, $q = 2.92$, ($p_1 = 0.02$), $\omega = 1.0$	0.10	−23434.07	M7 versus M8: 18.18, d.f. = 2, $P = .0001$	See Table 6

Table 5: Positive selection sites in the glycoprotein gene predicted by using Bayes empirical analysis under different PAML models.

Codon		Amino acid		Posterior probability		Post mean ± S.E.	
Dataset I	Dataset II	Dataset I	Dataset II	Dataset I	Dataset II	Dataset I	Dataset II
466	466	A	A	0.68	0.72	1.27 ± 0.35	1.29 ± 0.34
483	483	V	V	0.95	0.84	1.46 ± 0.16	1.39 ± 0.26
486	486	T	T	0.56	0.53	1.19 ± 0.36	1.16 ± 0.36
490	490	Q	Q	0.82	0.80	1.38 ± 0.27	1.36 ± 0.29

Dataset I: The whole 53 nucleotide sequences. Dataset II: AY987478 was excluded.

adaptation is due to the G gene. Positive selection is an important evolutionary force that drives adaptation. It is not surprising that evolutionary scientists first applied selection analysis to the G gene of lyssaviruses [39]. One notable difference between the previous investigations and our study was the dataset. Previous dataset with 55 complete G gene sequences were from isolates of natural rabies infections, excluding passages and vaccine strains. Our dataset included street 53 rabies isolates and vaccine strains collected over a period of 70 years from 21 countries. The neutrality tests on the G in lyssavirus indicated that the protein was under negative selection. Analysis of heterogeneous selective pressures on the amino acid sites across the gene found no evidence for positive selection on any site when the putative recombinant AY987478 was excluded. Instead, most of the sites were under strong negative selection, which was consistent with previous investigations using only street rabies isolates [39, 40]. The only weak positive selection identified by our analyses was at amino acid residue 483 (not in the ectodomain). No positive selection has been detected in the main epitope II or III, the site of virus escape identified by monoclonal antibody binding selections in vitro. It is possible that the results were confounded by the sequences from isolates under intensive cell culture. Repeated passages of an RNA virus resulted in loss of fitness due to Muller's ratchet [41]. Serial virus passages severely reduce population size when a small set of founder population is reintroduced into an identical unpopulated environment, which may lead to the stochastic loss of certain genotypes, especially the rare genotypes [42, 43]. However, exclusion of sequences of passaged lyssaviruses from the dataset in this study did not affect the readout of the analyses. It appears that rabies spillover, host shift (happened naturally), virus escape by monoclonal antibody selection, and vaccine strains (under various in vitro and in vivo conditions) is not the result of positive selection in the G gene.

Recombination is another important evolutionary driving force in adaptation, and it is a mechanism that prevents the accumulation of deleterious substitutions [44]. It allows the acquisition of multiple genetic changes in a single step and can combine genetic information to produce advantageous genotypes. It may be important for incremental host adaptation after switching to new host has occurred [45]. Recombination in rabies viruses had been proposed, but it was not thoroughly inspected [46, 47]. Our study suggested one recombinant event. The recombinant sequence AY987478 was from a dog isolate (CHAND03, genotype 1) and the possible parental sequences were isolated from dogs and sheep from the same geographic area (India and Nepal). However, the putative recombinant AY987478 could be an artifact from sequencing or sample contamination. Generation of recombinants in the course of reverse transcription of RNA and subsequent PCR is a well-known phenomenon [48–50]. From the bootscanning analysis in this research, the 3 prime and 5 prime regions of AY987478 were clustered with putative parents with a bootstrap value of 100%, indicating little difference between the two sequences in the two regions. By checking the sequences, there are regions of about 450 bases long that are identical between the recombinant and the corresponding parent, which is

Table 6: Detection of selection pressure on glycoprotein gene using methods implemented in the Datamonkey website.

Dataset	Mean dN/dS			Positive selection sites			Negative selection sites			Codon (P-Value)
	SLAC	FEL	REL	SLAC	FEL	REL	SLAC	FEL	REL	
Dataset I	0.1226		0.1278	0	0	0	397	418	0	
Dataset II	0.1231		0.1274	0	0	0	391	417	0	
Dataset III	0.1214		0.1233	0	1	0	386	416	0	416 (.0999)

Dataset I: The whole 53 nucleotide sequences. Dataset II: AY987478 was excluded. Dataset III: the six putative recombinants were excluded.

rare considering the high mutation rate in RNA viruses. The homologous recombination rate in negative-sense RNA virus was found to be low [46], which is supported by a recent report that homologous recombination is very rare or absent in influenza A virus [17]. Further experimentation is needed to prove that the recombinant AY987478 is not an artifact.

In summary, we did not find significant support for positive selection pressure on G gene in lyssavirus isolates from different rabies hosts and vaccine strains that cover 70 years of evolution in 21 countries. The recombination analysis suggested an orphan event that needs further investigation. It appears that evolution of the G gene may not play a major role in lyssavirus adaptation. It is surprising considering the functions of glycoprotein in lyssavirus infection. It has been reported that host switching from chiropters to carnivores has occurred in lyssavirus evolution history [7, 9]. Spillovers of lyssaviruses from chiropters to other animals may have happened repeatedly and still occur [8]. Transmission of European bt lyssavirus 1 (EBLV-1) was reported in sheep [51], stone marten [52], and cats [53]. For a successful spillover and subsequent adaptation, there must be effective cross-species viral exposure and compatibility between the virus and the new host to allow replication and transmission. Lyssavirus infections are typically transmitted by the virus-laden saliva of a rabid animal via a bite or scratch, which can facilitate cross-species viral exposures. The initial viral interaction with cells of a new host plays a critical role in determining host specificity and host shift [45]. For example, feline virus acquired the ability to infect dogs through changes in its capsid protein that binds to canine transferrin receptor on canine cells [54]. Lyssavirus G is a surface glycoprotein responsible for receptor recognition and membrane fusion [7–9, 55]. It is reasonable to expect that the protein is under positive selection pressure in the viral adaptation to the new host. The lack of positive selection in the G glycoprotein suggests that the virus is not subject to strong immune selection [25]. The G gene may escape the immunity of the host since lyssaviruses migrate from the peripheral to the central nervous systems [7]. Recent investigation demonstrated that diminishing frequencies of both cross-species transmission and host shifts were found with increasing phylogenetic distance between bat species [9], indicating the virus, thus the G gene, is subject to less selection pressure in a similar host and cellular environment [7, 25]. However, the G gene might have been under relative low positive selection that was not detected by current computational methods. More sensitive method or properly relaxed statistical significance stringency with experimental verification may help identify the role of the G gene in lyssavirus adaptation.

Acknowledgments

The authors thank Jan Pohl, Elizabeth Neuhaus, and Charles E. Rupprecht for support in this investigation. They also thank Kathryn Kellar and Scott Sammons for helpful suggestions to the paper. Use of trade names and commercial sources are for identification only and do not imply endorsement by the U S Department of Health and Human Services. The findings and conclusions in this paper are those of the authors and do not necessarily represent the views of the funding agency.

References

[1] T. Lefébure and M. J. Stanhope, “Evolution of the core and pan-genome of Streptococcus: positive selection, recombination, and genome composition,” *Genome Biology*, vol. 8, no. 5, pp. R71.1–R71.16, 2007.

[2] D. A. Steinhauer, E. Domingo, and J. J. Holland, “Lack of evidence for proofreading mechanisms associated with an RNA virus polymerase,” *Gene*, vol. 122, no. 2, pp. 281–288, 1992.

[3] K. Kirkegaard and D. Baltimore, “The mechanism of RNA recombination in poliovirus,” *Cell*, vol. 47, no. 3, pp. 433–443, 1986.

[4] M. Worobey and E. C. Holmes, “Evolutionary aspects of recombination in RNA viruses,” *Journal of General Virology*, vol. 80, no. 10, pp. 2535–2543, 1999.

[5] D. M. Pfukenyi, D. Pawandiwa, P. V. Makaya, and U. Ushewokunze-Obatolu, “A retrospective study of wildlife rabies in Zimbabwe,” *Tropical Animal Health and Production*, vol. 41, no. 4, pp. 565–572, 2009.

[6] A. I. Wandeler, S. A. Nadin-Davis, R. R. Tinline, and C. E. Rupprecht, “Rabies epidemiology: some ecological and evolutionary perspectives,” in *Lyssaviruses*, C. E. Rupprecht, B. Dietzchold, and H. Koprowski, Eds., pp. 297–324, Springer, Berlin, Germany, 1994.

[7] H. Badrane and N. Tordo, “Host switching in Lyssavirus history from the chiroptera to the carnivora orders,” *Journal of Virology*, vol. 75, no. 17, pp. 8096–8104, 2001.

[8] L. K. Crawford-Miksza, D. A. Wadford, and D. P. Schnurr, “Molecular epidemiology of enzootic rabies in California,” *Journal of Clinical Virology*, vol. 14, no. 3, pp. 207–219, 1999.

[9] D. G. Streicker, A. S. Turmelle, M. J. Vonhof, I. V. Kuzmin, G. F. McCracken, and C. E. Rupprecht, “Host phylogeny constrains cross-species emergence and establishment of rabies virus in bats,” *Science*, vol. 329, no. 5992, pp. 676–679, 2010.

[10] B. Dietzschold, W. H. Wunner, T. J. Wiktor et al., "Characterization of an antigenic determinant of the glycoprotein that correlates with pathogenicity of rabies virus," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 80, no. 1, pp. 70–74, 1983.

[11] X. Yan, P. S. Mohankumar, B. Dietzschold, M. J. Schnell, and Z. F. Fu, "The rabies virus glycoprotein determines the distribution of different rabies virus strains in the brain," *Journal of Neurovirology*, vol. 8, no. 4, pp. 345–352, 2002.

[12] R. K. Bradley, A. Roberts, M. Smoot et al., "Fast statistical alignment," *PLos Computational Biology*, vol. 5, no. 5, Article ID e1000392, 2009.

[13] K. Tamura, J. Dudley, M. Nei, and S. Kumar, "MEGA4: molecular evolutionary genetics analysis (MEGA) software version 4.0," *Molecular Biology and Evolution*, vol. 24, no. 8, pp. 1596–1599, 2007.

[14] T. C. Bruen, H. Philippe, and D. Bryant, "A simple and robust statistical test for detecting the presence of recombination," *Genetics*, vol. 172, no. 4, pp. 2665–2681, 2006.

[15] I. B. Jakobsen and S. Easteal, "A program for calculating and displaying compatibility matrices as an aid in determining reticulate evolution in molecular sequences," *Computer Applications in the Biosciences*, vol. 12, no. 4, pp. 291–295, 1996.

[16] J. M. Smith, "Analyzing the mosaic structure of genes," *Journal of Molecular Evolution*, vol. 34, no. 2, pp. 126–129, 1992.

[17] M. F. Boni, D. Posada, and M. W. Feldman, "An exact nonparametric method for inferring mosaic structure in sequence triplets," *Genetics*, vol. 176, no. 2, pp. 1035–1047, 2007.

[18] S. L. Kosakovsky Pond, D. Posada, M. B. Gravenor, C. H. Woelk, and S. D. W. Frost, "GARD: a genetic algorithm for recombination detection," *Bioinformatics*, vol. 22, no. 24, pp. 3096–3098, 2006.

[19] S. L. Kosakovsky Pond and S. D. W. Frost, "Datamonkey: rapid detection of selective pressure on individual sites of codon alignments," *Bioinformatics*, vol. 21, no. 10, pp. 2531–2533, 2005.

[20] K. S. Lole, R. C. Bollinger, R. S. Paranjape et al., "Full-length human immunodeficiency virus type 1 genomes from subtype C- infected seroconverters in India, with evidence of intersubtype recombination," *Journal of Virology*, vol. 73, no. 1, pp. 152–160, 1999.

[21] Z. Yang, "PAML 4: phylogenetic analysis by maximum likelihood," *Molecular Biology and Evolution*, vol. 24, no. 8, pp. 1586–1591, 2007.

[22] Z. Yang, R. Nielsen, N. Goldman, and A. M. K. Pedersen, "Codon-substitution models for heterogeneous selection pressure at amino acid sites," *Genetics*, vol. 155, no. 1, pp. 431–449, 2000.

[23] Z. Yang, W. S. W. Wong, and R. Nielsen, "Bayes empirical Bayes inference of amino acid sites under positive selection," *Molecular Biology and Evolution*, vol. 22, no. 4, pp. 1107–1118, 2005.

[24] H. Badrane, C. Bahloul, P. Perrin, and N. Tordo, "Evidence of two Lyssavirus phylogroups with distinct pathogenicity and immunogenicity," *Journal of Virology*, vol. 75, no. 17, pp. 3268–3276, 2001.

[25] K. J. Guyatt, J. Twin, P. Davis et al., "A molecular epidemiology study of Australian bat lyssavirus," *Journal of General Virology*, vol. 84, no. 2, pp. 485–496, 2003.

[26] Q. Tang, L. A. Orciari, C.E. Rupprecht, and X. Zhao, "Sequencing and positional analysis of the glycoprotein gene of four Chinese rabies viruses," *Zhongguo Bingduxue*, vol. 15, no. 1, pp. 22–33, 2000 (Chinese).

[27] T. Hemachudha, S. Wacharapluesadee, B. Lumlertdaecha et al., "Sequence analysis of rabies virus in humans exhibiting encephalitic or paralytic rabies," *Journal of Infectious Diseases*, vol. 188, no. 7, pp. 960–966, 2003.

[28] H. Ito, N. Minamoto, T. Watanabe et al., "A unique mutation of glycoprotein gene of the attenuated RC-HL strain of rabies virus, a seed virus used for production of animal vaccine in Japan," *Microbiology and Immunology*, vol. 38, no. 6, pp. 479–482, 1994.

[29] S. Agrawal, A. K. Shasany, and S. P. S. Khanuja, "Plant transformation vectors having street rabies virus (Indian strain) glycoprotein gene," *Journal of Plant Biochemistry and Biotechnology*, vol. 14, no. 2, pp. 81–87, 2005.

[30] B. H. Hyun, K. K. Lee, IN. J. Kim et al., "Molecular epidemiology of rabies virus isolates from South Korea," *Virus Research*, vol. 114, no. 1-2, pp. 113–125, 2005.

[31] S. L. Meng, J. X. Yan, GE. L. Xu et al., "A molecular epidemiological study targeting the glycoprotein gene of rabies virus isolates from China," *Virus Research*, vol. 124, no. 1-2, pp. 125–138, 2007.

[32] X. Bai, C. K. Warner, and M. Fekadu, "Comparisons of nucleotide and deduced amino acid sequences of the glycoprotein genes of a Chinese street strain (CGX89-1) and a Chinese vaccine strain (3aG) of rabies virus," *Virus Research*, vol. 27, no. 2, pp. 101–112, 1993.

[33] E. Yelverton, S. Norton, J. F. Obijeski, and D. V. Goeddel, "Rabies virus glycoprotein analogs: biosynthesis in Escherichia coli," *Science*, vol. 219, no. 4585, pp. 614–619, 1983.

[34] A. Benmansour, M. Brahimi, C. Tuffereau, P. Coulon, F. Lafay, and A. Flamand, "Rapid sequence evolution of street rabies glycoprotein is related to the highly heterogeneous nature of the viral population," *Virology*, vol. 187, no. 1, pp. 33–45, 1992.

[35] S. A. Nadin-Davis, G. Allen Casey, and A. I. Wandeler, "A molecular epidemiological study of rabies virus in central Ontario and western Quebec," *Journal of General Virology*, vol. 75, no. 10, pp. 2575–2583, 1994.

[36] S. A. Nadin-Davis, M. I. Sampath, G. A. Casey, R. R. Tinline, and A. I. Wandeler, "Phylogeographic patterns exhibited by Ontario rabies virus variants," *Epidemiology and Infection*, vol. 123, no. 2, pp. 325–336, 1999.

[37] S. A. Nadin-Davis, W. Huang, and A. I. Wandeler, "The design of strain-specific polymerase chain reactions for discrimination of the racoon rabies virus strain fron indigenous rabies viruses of Ontario," *Journal of Virological Methods*, vol. 57, no. 1, pp. 1–14, 1996.

[38] K. Morimoto, M. Patel, S. Corisdeo et al., "Characterization of a unique variant of bat rabies virus responsible for newly emerging human cases in North America," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 93, no. 11, pp. 5653–5658, 1996.

[39] E. C. Holmes, C. H. Woelk, R. Kassis, and H. Bourhy, "Genetic constraints and the adaptive evolution of rabies virus in nature," *Virology*, vol. 292, no. 2, pp. 247–257, 2002.

[40] H. Bourhy, J. M. Reynes, E. J. Dunham et al., "The origin and phylogeography of dog rabies virus," *Journal of General Virology*, vol. 89, no. 11, pp. 2673–2681, 2008.

[41] D. K. Clarke, E. A. Duarte, A. Moya, S. F. Elena, E. Domingo, and J. Holland, "Genetic bottlenecks and population passages cause profound fitness differences in RNA viruses," *Journal of Virology*, vol. 67, no. 1, pp. 222–228, 1993.

[42] M. Oberle, O. Balmer, R. Brun, and I. Roditi, "Bottlenecks and the maintenance of minor genotypes during the life cycle of Trypanosoma brucei," *PLos Pathogens*, vol. 6, no. 7, Article ID e1001023, 2010.

[43] L. M. Wahl, P. J. Gerrish, and I. Saika-Voivod, "Evaluating the impact of population bottlenecks in experimental evolution," *Genetics*, vol. 162, no. 2, pp. 961–971, 2002.

[44] M. Poss, A. Idoine, H. A. Ross, J. A. Terwee, S. VandeWoude, and A. Rodrigo, "Recombination in feline lentiviral genomes during experimental cross-species infection," *Virology*, vol. 359, no. 1, pp. 146–151, 2007.

[45] C. R. Parrish, E. C. Holmes, D. M. Morens et al., "Cross-species virus transmission and the emergence of new epidemic diseases," *Microbiology and Molecular Biology Reviews*, vol. 72, no. 3, pp. 457–470, 2008.

[46] E. R. Chare, E. A. Gould, and E. C. Holmes, "Phylogenetic analysis reveals a low rate of homologous recombination in negative-sense RNA viruses," *Journal of General Virology*, vol. 84, no. 10, pp. 2691–2703, 2003.

[47] L. Geue, S. Schares, C. Schnick et al., "Genetic characterisation of attenuated SAD rabies virus strains used for oral vaccination of wildlife," *Vaccine*, vol. 26, no. 26, pp. 3227–3235, 2008.

[48] A. Flockerzi, J. Maydt, O. Frank et al., "Expression pattern analysis of transcribed HERV sequences is complicated by ex vivo recombination," *Retrovirology*, vol. 4, no. 39, pp. 1–12, 2007.

[49] G. Luo and J. Taylor, "Template switching by reverse transcriptase during DNA synthesis," *Journal of Virology*, vol. 64, no. 9, pp. 4321–4328, 1990.

[50] A. Meyerhans, J. P. Vartanian, and S. Wain-Hobson, "DNA recombination during PCR," *Nucleic Acids Research*, vol. 18, no. 7, pp. 1687–1691, 1990.

[51] H. Bourhy, L. Dacheux, C. Strady, and A. Mailles, "Rabies in Europe in 2005," *Euro Surveillanc*, vol. 10, no. 11, pp. 213–216, 2005.

[52] K. Tjørnehøj, A. R. Fooks, J. S. Agerholm, and L. Rønsholt, "Natural and experimental infection of sheep with european bat lyssavirus type-1 of danish bat origin," *Journal of Comparative Pathology*, vol. 134, no. 2-3, pp. 190–201, 2006.

[53] L. Dacheux, F. Larrous, A. Mailles et al., "European bat lyssavirus transmission among cats, Europe," *Emerging Infectious Diseases*, vol. 15, no. 2, pp. 280–284, 2009.

[54] K. Hueffer, J. S. L. Parker, W. S. Weichert, R. E. Geisel, J. Y. Sgro, and C. R. Parrish, "The natural host range shift and subsequent evolution of canine parvovirus resulted from virus-specific binding to the canine transferrin receptor," *Journal of Virology*, vol. 77, no. 3, pp. 1718–1726, 2003.

[55] P. Durrer, Y. Gaudin, R. W. H. Ruigrok, R. Graf, and J. Brunner, "Photolabeling identifies a putative fusion domain in the envelope glycoprotein of rabies and vesicular stomatitis viruses," *Journal of Biological Chemistry*, vol. 270, no. 29, pp. 17575–17581, 1995.

SAGE-Hindawi Access to Research
International Journal of Evolutionary Biology
Volume 2011, Article ID 379424, 15 pages
doi:10.4061/2011/379424

Review Article

Baculovirus: Molecular Insights on Their Diversity and Conservation

Solange Ana Belen Miele, Matías Javier Garavaglia, Mariano Nicolás Belaich, and Pablo Daniel Ghiringhelli

LIGBCM (Laboratorio de Ingeniería Genética y Biología Celular y Molecular), Departamento de Ciencia y Tecnología, Universidad Nacional de Quilmes, Roque Saenz Peña 352, Bernal, Argentina

Correspondence should be addressed to Pablo Daniel Ghiringhelli, pdghiringhelli@gmail.com

Received 15 October 2010; Revised 4 January 2011; Accepted 14 February 2011

Academic Editor: Kenro Oshima

The Baculoviridae is a large group of insect viruses containing circular double-stranded DNA genomes of 80 to 180 kbp. In this study, genome sequences from 57 baculoviruses were analyzed to reevaluate the number and identity of core genes and to understand the distribution of the remaining coding sequences. Thirty one core genes with orthologs in all genomes were identified along with other 895 genes differing in their degrees of representation among reported genomes. Many of these latter genes are common to well-defined lineages, whereas others are unique to one or a few of the viruses. Phylogenetic analyses based on core gene sequences and the gene composition of the genomes supported the current division of the Baculoviridae into 4 genera: *Alphabaculovirus, Betabaculovirus, Gammabaculovirus*, and *Deltabaculovirus*.

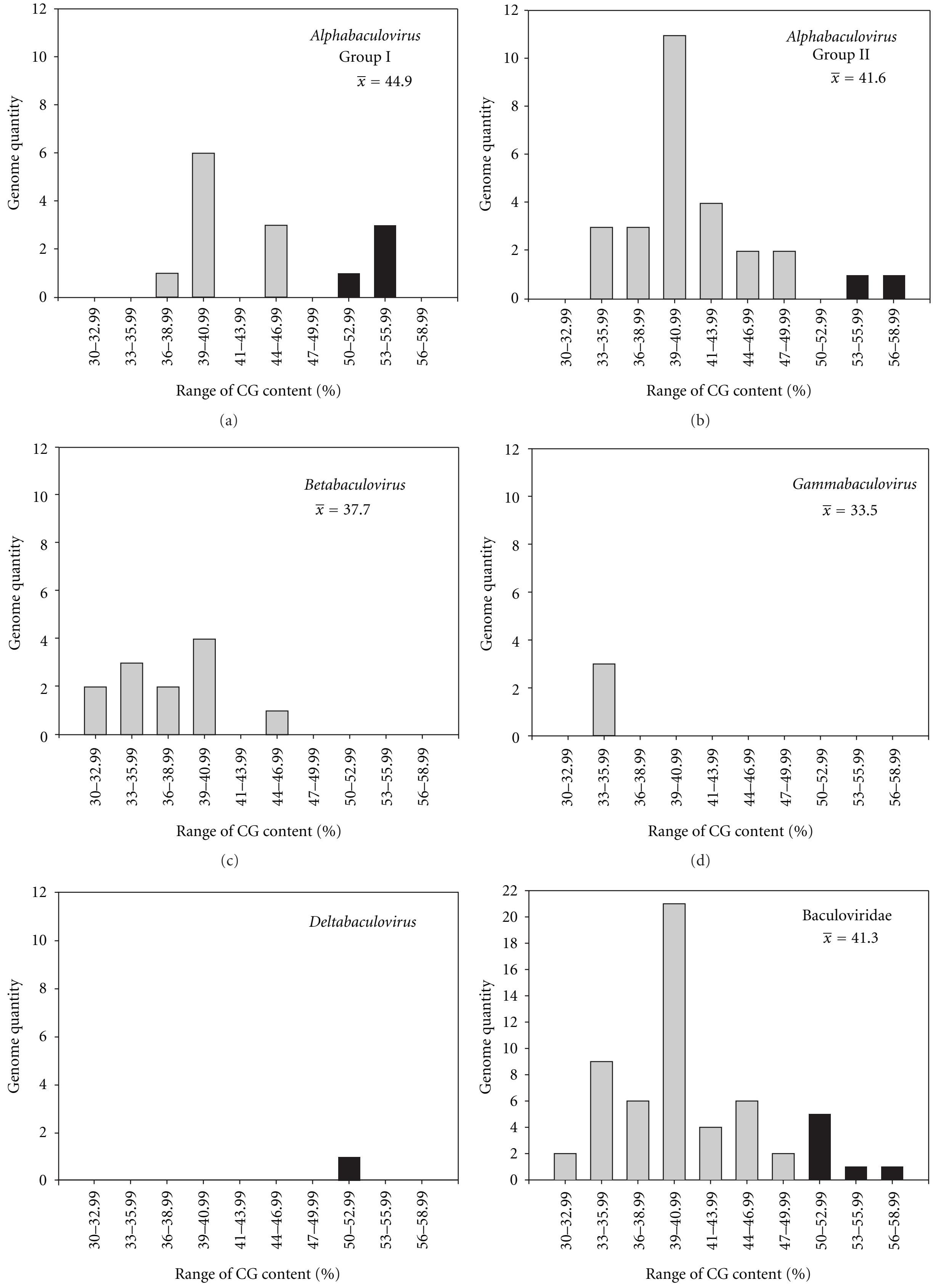
1. Background

Baculoviruses are arthropod-specific viruses containing large double-stranded circular DNA genomes of 80,000–180,000 bp. The progeny generation is biphasic, with two different phenotypes during virus infection: budded viruses (BVs), during the initial stage of the multiplication cycle, and occlusion-derived viruses (ODVs), at the final stages of replication [1, 2]. In general, primary infection takes place in the insect midgut cells after ingestion of occlusion bodies (OBs). Following this stage, systemic infection is caused by the initial BV progeny [3, 4]. And finally, OBs are produced during the last stage of the infection. These OBs comprise virions embedded in a protein matrix which protects them from the environment [5, 6].

Baculoviruses have been used extensively in many biological applications such as protein expression systems, models of genetic regulatory networks and genome evolution, putative nonhuman viral vectors for gene delivery, and biological control agents against insect pests [7–17].

The Baculoviridae family is divided into four genera according to common biological and structural characteristics: *Alphabaculovirus*, which includes lepidopteran-specific baculoviruses and is subdivided into Group I or Group II based on the type of fusogenic protein, *Betabaculovirus*, comprising lepidopteran-specific granuloviruses, *Gammabaculovirus*, which includes hymenopteran-specific baculoviruses, and finally *Deltabaculovirus* which, to date, comprises only CuniNPV and possibly the still undescribed dipteran-specific baculoviruses [1, 18–20].

The comparison between known genome sequences of all baculoviruses has been the source for identifying a common set of genes, the baculovirus core genes. However, there are probably more orthologous sequences that may not be identified due to the accumulation of many mutations throughout evolution. Thus, core genes seem to be a key factor for some of the main biological functions, such as those necessary to transcribe viral late genes, produce virion structure, infect gut cells abrogate host metabolism and establish infections [21–24].



(a) (b) (c) (d) (e) (f)

Figure 1: GC content in baculovirus genomes. The different histograms contain the distribution of baculovirus genomes according to their GC content and their genus classification. Black bars highlight genomes with a GC content higher than 50%.

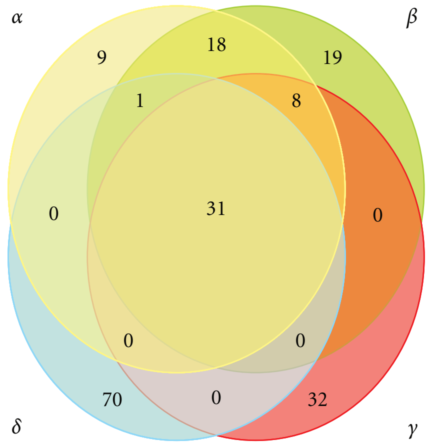


Figure 2: Baculovirus core genes. The different circles represent the 4 baculovirus genera (in yellow *Alphabaculovirus*; in green *Betabaculovirus*; in red *Gammabaculovirus*; in blue *Deltabaculovirus*). The numbers contained within the overlapping regions indicate the amount of shared genes between all members of the genera. The numbers within the circles but outside the overlapping regions indicate the amount of genes shared by all members of that genus but with the absence of orthologous sequences in the remaining genera. These estimations were inferred by Blast P algorithm (http://www.ncbi.nlm.nih.gov/) considering $E = 0.001$ as cutoff value and comparing all reported baculovirus ORFs between them. The identity of common genes is provided in the Supplementary data available at doi:10.4061/2011/379424

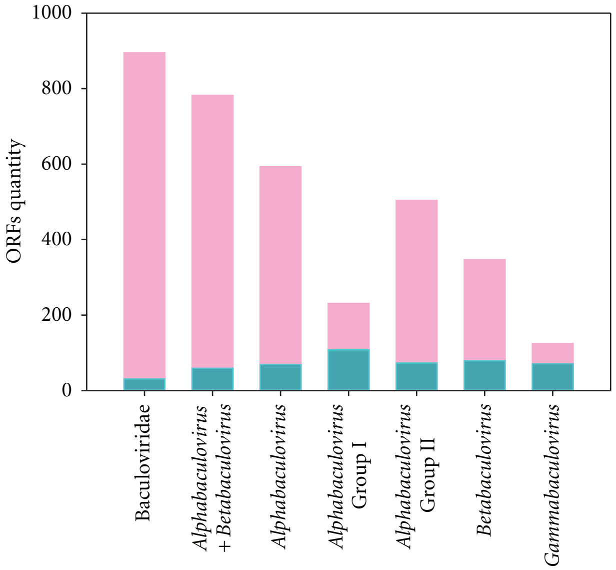


Figure 3: Whole baculovirus gene content. The histogram shows the amount of different reported genes in each baculovirus genus or recognized lineage (bars in pink color), and the subset of shared genes for all members of the corresponding phylogenetic clade (bars in green color). This bar graph was performed using the information resulting from the comparison of all ORFs reported in the 57 baculovirus with known genomes, analyzing all against all by Blast P algorithm (http://www.ncbi.nlm.nih.gov/) considering $E = 0.001$ as cutoff value.

For this report, previous data as well as bioinformatic studies conducted on currently available sets of completely sequenced baculovirus genomes were taken into account and have resulted in a summary of gene content and phylogenetic analyses which validates the classification of this important viral family.

2. Baculovirus Ancestral Genes

There are currently 57 complete baculovirus genomes deposited in GenBank (Table 1). These include 41 *Alphabaculoviruses*, 12 *Betabaculoviruses*, 3 *Gammabaculoviruses*, and 1 *Deltabaculovirus*.

As a first approach to perform a comparative analysis, the GC content of the genomes were calculated (Figure 1). The histogram revealed that many baculoviruses have about 41% of GC content although several of them have significantly higher values (CfMNPV at 50.1%, CuniNPV at 50.9%, AnpeNPV-L2 at 53.5%, AnpeNPV-Z at 53.5%, LyxyNPV at 53.5%, OpMNPV at 55.1%, and LdMNPV at 57.5%). A detailed analysis of DNA content did not show a clear pattern of GC content that could be associated with each genus.

Further characterization of the patterns of gene content and organization may prove useful for establishing evolutionary relationships among members of Baculoviridae. The high variability observed in the number of coding sequences becomes a key feature of viruses with large DNA genomes that infect eukaryotic cells [18]. Insertions, deletions, duplication events, and/or sequence reorganizations by recombination or transposition processes seem to be the main forces of the macroevolution in this particular kind of biological entities. For example, the loss or gain of genetic material could provide new important abilities for colonization of new hosts, or they could improve performance within established hosts. However, there seems to be a set of core genes whose absence would imply the loss of basic biological functions, and that could be typical of the viral family. In view of this, and considering previous reports [1, 19, 22, 23], the amount and identity of baculovirus common genes were reevaluated (Table 2). As a result, P6.9 and Desmoplakin were recognized in this work, as core proteins by using sequence analysis complementary to the standard ones (see Supplementary files available at doi:10.4061/2011/379424).

The group of conserved sequences found in all baculovirus genomes is consistently estimated at about 30 shared genes, regardless of the increasing number of genomes analyzed [22, 148]. Meanwhile, the role or function assigned to several sequences has been renewed, according to new studies. In particular, it has been identified that 38k (Ac98) gene encodes a protein which is part of the capsid structure [121, 122]; P33 (Ac92) is a sulfhydryl oxidase which could be related to the proper production of virions in the infected cell nucleus [123–125]; ODV-EC43 (Ac109) is a structural component which would be involved in BV and ODV generation [126]; P49 (Ac142) is a capsid

Table 1: Baculovirus complete genomes.

Genus	Name	Abbreviation	Code	Accesion number	Genome (bp)	Annotated ORFs	GC%	Ref.
Alphabaculovirus-Group I	*Antheraea pernyi* NPV-Z	AnpeNPV-Z	APN	NC_008035	126629	145	53.5	[27]
	Antheraea pernyi NPV-L2	AnpeNPV-L2	AP2	EF207986	126246	144	53.5	[28]
	Anticarsia gemmatalis MNPV-2D	AgMNPV-2D	AGN	NC_008520	132239	152	44.5	[29]
	Autographa californica MNPV-C6	AcMNPV-C6	ACN	NC_001623	133894	154	40.7	[30]
	Bombyx mori NPV	BmNPV	BMN	NC_001962	128413	137	40.4	[31]
	Bombyx mandarina NPV	BomaNPV	BON	NC_012672	126770	141	40.2	[32]
	Choristoneura fumiferana DEF MNPV	CfDEFMNPV	CDN	NC_005137	131160	149	45.8	[33]
	Choristoneura fumiferana MNPV	CfMNPV	CFN	NC_004778	129593	145	50.1	[34]
	Epiphyas postvittana NPV	EppoNPV	EPN	NC_003083	118584	136	40.7	[35]
	Hyphantria cunea NPV	HycuNPV	HCN	NC_007767	132959	148	45.5	[36]
	Maruca vitrata MNPV	MaviMNPV	MVN	NC_008725	111953	126	38.6	[37]
	Orgyia pseudotsugata MNPV	OpMNPV	OPN	NC_001875	131995	152	55.1	[38]
	Plutella xylostella MNPV	PlxyMNPV	PXN	NC_008349	134417	149	40.7	U
	Rachiplusia ou MNPV	RoMNPV	RON	NC_004323	131526	146	39.1	[39]
Alphabaculovirus-Group II	*Adoxophyes honmai* NPV	AdhoNPV	AHN	NC_004690	113220	125	35.6	[40]
	Adoxophyes orana NPV	AdorNPV	AON	NC_011423	111724	121	35.0	[41]
	Agrotis ipsilon NPV	AgipNPV	AIN	NC_011345	155122	163	48.6	U
	Agrotis segetum NPV	AgseNPV	ASN	NC_007921	147544	153	45.7	[42]
	Apocheima cinerarium NPV	ApciNPV	APO	FJ914221	123876	118	33.4	U
	Chrysodeixis chalcites NPV	ChChNPV	CCN	NC_007151	149622	151	39.0	[43]
	Clanis bilineata NPV	ClbiNPV	CBN	NC_008293	135454	129	37.7	[44]
	Ectropis obliqua NPV	EcobNPV	EON	NC_008586	131204	126	37.6	[45]
	Euproctis pseudoconspersa NPV	EupsNPV	EUN	NC_012639	141291	139	40.4	[46]
	Helicoverpa armigera NPV-C1	HearNPV-C1	HA1	NC_003094	130759	135	38.9	[47]
	Helicoverpa armigera NPV-G4	HearNPV-G4	HA4	NC_002654	131405	135	39.0	[47]
	Helicoverpa armigera MNPV	HearMNPV	HAN	NC_011615	154196	162	40.1	[48]
	Helicoverpa armigera SNPV-NNg1	HearSNPV-NNg1	HAS	NC_011354	132425	143	39.2	[49]
	Helicoverpa zea SNPV	HzSNPV	HZN	NC_003349	130869	139	39.1	U
	Leucania separata NPV-AH1	LeseNPV-AH1	LSN	NC_008348	168041	169	48.6	[50]
	Lymantria dispar MNPV	LdMNPV	LDN	NC_001973	161046	163	57.5	[51]
	Lymantria xylina MNPV	LyxyMNPV	LXN	NC_013953	156344	157	53.5	[52]

Table 1: Continued.

Genus	Name	Abbreviation	Code	Accesion number	Genome (bp)	Annotated ORFs	GC%	Ref.
	Mamestra configurata NPV-90-2	MacoNPV-90-2	MCN	NC_003529	155060	169	41.7	[53]
	Mamestra configurata NPV-90-4	MacoNPV-90-4	MC4	AF539999	153656	168	41.7	[54]
	Mamestra configurata NPV-B	MacoNPV-B	MCB	NC_004117	158482	169	40.0	[55]
	Orgyia leucostigma NPV	OrleNPV	OLN	NC_010276	156179	135	39.9	U
	Spodoptera exigua MNPV	SeMNPV	SEN	NC_002169	135611	142	43.8	U
	Spodoptera frugiperda MNPV-3AP2	SfMNPV-3AP2	SF2	NC_009011	131330	143	40.2	[56]
	Spodoptera frugiperda MNPV-19	SfMNPV-19	SF9	EU258200	132565	141	40.3	[57]
	Spodoptera litura NPV-II	SpliNPV-II	SLN	NC_011616	148634	147	45.0	U
	Spodoptera litura NPV-G2	SpliNPV-G2	SL2	NC_003102	139342	141	42.8	[58]
	Trichoplusia ni SNPV	TnSNPV	TNN	NC_007383	134394	144	39.0	[59]
Betabaculovirus	*Adoxophyes orana* GV	AdorGV	AOG	NC_005038	99657	119	34.5	[60]
	Agrotis segetum GV	AgseGV	ASG	NC_005839	131680	132	37.3	U
	Choristoneura occidentalis GV	ChocGV	COG	NC_008168	104710	116	32.7	[61]
	Cryptophlebia leucotreta GV	CrleGV	CLG	NC_005068	110907	129	32.4	[62]
	Cydia pomonella GV	CpGV	CPG	NC_002816	123500	143	45.3	[63]
	Helicoverpa armigera GV	HearGV	HAG	NC_010240	169794	179	40.8	[64]
	Phthorimea operculella GV	PhopGV	POG	NC_004062	119217	130	35.7	[65]
	Plutella xylostella GV	PlxyGV	PXG	NC_002593	100999	120	40.7	[66]
	Pieris rapae GV	PiraGV	PRG	GQ884143	108592	120	33.2	U
	Pseudaletia unipuncta GV-Hawaiin	PsunGV	PUG	EU678671	176677	183	39.8	U
	Spodoptera litura GV-K1	SpliGV	SLG	NC_009503	124121	136	38.8	[67]
	Xestia c-nigrum GV	XnGV	XCG	NC_002331	178733	181	40.7	[68]
Gamma	*Neodiprion abietis* NPV	NeabNPV	NAN	NC_008252	84264	93	33.4	[69]
	Neodiprion lecontei NPV	NeleNPV	NLN	NC_005906	81755	93	33.3	[70, 71]
	Neodiprion sertifer NPV	NeseNPV	NSN	NC_005905	86462	90	33.8	[71, 72]
Delta	*Culex nigripalpus* NPV	CuniNPV	CNN	NC_003084	108252	109	50.9	[73]

This table contains all of baculoviruses used in bioinformatic studies, sorted by genus (and within them by alphabetical order). MNPV is the abbreviation of multicapsid nucleopolyhedrovirus; NPV is the abbreviation of nucleopolyhedrovirus; SNPV is the abbreviation of single nucleopolyhedrovirus; GV is the abbreviation of granulovirus. The accession numbers are from National Center for Biotechnology Information (NCBI, http://www.ncbi.nlm.nih.gov/) and correspond to the sequences of complete genomes. Code is an acronym used for practicality. U: unpublished.

protein important in DNA processing, packaging, and capsid morphogenesis [129]; Ac81 interacts with Actin 3 in the cytoplasm but does not appear in BVs or in ODVs [135]; ODV-E18 (Ac143) would mediate BV production [131]; desmoplakin (Ac66) seems to be essential in releasing processes from virogenic stroma to cytoplasm [132]; PIF-4 (Ac96) and PIF-5 (ODV-56, Ac148) are ODV envelope proteins with an essential role in *per os* infection route [145, 147]; Ac68 may be involved in polyhedron morphogenesis [130].

Table 2: Core genes.

	ACN	LDN	CPG	NSN	CNN
Replication					
lef-1 [74]	14	123	74	68	45
lef-2 [74]	6	137	41	57	25
DNA pol [75–78]	65	83	111	28	91
Helicase [79–90]	95	97	90	61	89
Transcription					
lef-4 [91–95]	90	93	95	62	96
lef-8 [91, 96]	50	51	131	81	26
lef-9 [95, 97]	62	64	117	40	59
p47 [91, 98]	40	48	68	49	73
lef-5 [98–101]	99	100	87	58	88
Packaging, assembly, and release					
p6.9 [102–104]	100	101	86	36	23
vp39 [105–108]	89	92	96	89	24
vlf-1 [100, 109–113]	77	86	106	45	18
alk-exo [114–116]	133	157	125	31	53
vp1054 [117]	54	57	138	85	8
vp91/p95 [118]	83	91	101	84	35
gp41 [119, 120]	80	88	104	47	33
38 k [121, 122]	98	99	88	59	87
p33 [123–125]	92	94	93	24	14
odv-ec43 [126–128]	109	107	55	70	69
p49 [129]	142	20	15	63	30
odv-nc42 [130]	68	80	114	41	58
odv-e18 [131]	143	19	14	65	31
desmoplakin [132]	66	82	112	29	92
Cell cycle arrest and/or interaction with host proteins					
odv-e27 [133, 134]	144	18	97	66	32
ac81 [135]	81	89	103	48	106
Oral infectivity					
pif-0/p74 [136–141]	138	27	60	50	74
pif-1 [142–144]	119	155	75	79	29
pif-2 [136, 142]	22	119	48	55	38
pif-3 [142]	115	143	35	69	46
pif-4/19k/odv-e28 [145]	96	98	89	60	90
pif-5/odv-e56 [146, 147]	148	14	18	38	102

The virus names are indicated in three letter code according to established in Table 1.
Numbers in columns indicates the corresponding ORFs of each genome.

The number and identity of shared orthologous genes in every accepted member of each genus were investigated, and the unique sequences typical of each clade as well as those shared between different phylogenetic groups were identified (Figure 2).

This analysis shows that the four accepted baculovirus genera have accumulated a large number of genes during evolution. Probably, many of these sequences have been incorporated into viral genomes prior to diversification processes since they are found in members of different genera. In contrast, other genes are unique to each genus, suggesting that they have been incorporated more recently and after diversification (Table 3). The possibility that non-shared genes found only in one genus which represent baculovirus ancestral sequences deleted in the other lineages should also be considered. In any case, a set of particular genes which could help in an appropriate genus taxonomy of new baculoviruses with partial sequence information were obtained from this analysis.

3. Whole Baculovirus Gene Content

The study of all genes reported in the 57 completely sequenced viral genomes revealed the existence of about

Table 3: Shared genes*.

Core genes	lef-2 (ACN6), lef-1 (ACN14), pif-2 (ACN22), p47 (ACN40), lef-8 (ACN50), vp1054 (ACN54), lef-9 (ACN62), DNA polymerase (ACN65), Desmoplakin (ACN66), ACN68, vlf-1 (ACN77), gp41 (ACN80), ACN81, vp91/p95 (ACN83), vp39 (ACN89), lef-4 (ACN90), p33 (ACN92), helicase (ACN95), 19K (ACN96), 38 K (ACN98), lef-5 (ACN99), p6.9 (ACN100), odv-ec43 (ACN109), PIF-3 (ACN115), pif-1 (ACN119), alkaline exonuclease (ACN133), p74 (ACN138), p49 (ACN142), odv-e18 (ACN143), odv-e27 (ACN144), odv-e56 (ACN148)
Alpha + Beta + Gamma	Polh (ACN8), dbp (ACN25), p48 (ACN103), ACN145, pp34/PEP (ACN131), odv-e25 (ACN94), p40 (ACN101), ACN106/107
Alpha + Beta + Delta	F-protein (ACN23)
Alpha + Beta	pk-1 (ACN10), 38,7 kDa (ACN13), lef-6 (ACN28), pp31/39K (ACN36), ACN38, ACN53, 25K FP (ACN61), LEF-3 (ACN67), ACN75, ACN76, tlp20 (ACN82), p18 (ACN93), P12 (ACN102), ACN108, p24 (ACN129), me53 (ACN139), ACN146, ie-1 (ACN147)
Alpha	orf1629 capsid (ACN9), ACN19, pkip-1 (ACN24), ACN34, ACN51, iap-2 (ACN58/59), ACN104, p87/vp80 (ACN141), ie-0 (ACN71)
Alpha Group I	ptp-1/bvp (ACN1), ACN5, odv-e26 (ACN16), iap-1 (ACN27), ACN30, ACN72, ACN73, ACN114, ACN124, gp64 (ACN128), p25 (ACN132), ie-2 (ACN151)
Beta	CPG4, CPG5, CPG20, CPG23, CPG29, CPG33, CPG39, CPG45, Metalloproteinase (CPG46), CPG62, FGF-1 (CPG76), CPG79, CPG99, CPG100, CPG115, IAP-5 (CPG116), CPG123, CPG135, FGF-3 (CPG140)
Gamma	NSN3, NSN9, NSN11, NSN12, NSN13, NSN16, NSN18, NSN19, NSN20, NSN26, NSN29, NSN34, NSN37, NSN39, NSN42, NSN43, NSN44, NSN51, NSN52, NSN53, NSN54, NSN56, NSN64, NSN72, NSN74, NSN76, NSN77, NSN79, NSN82, NSN85, NSN86, NSN89
Delta	CNN2, CNN3, CNN6, CNN7, CNN9, CNN10, CNN11, CNN12, CNN13, CNN15, CNN16, CNN17, CNN20, CNN21, CNN22, CNN27, CNN28, CNN31, CNN36, CNN37, CNN39, CNN40, CNN41, CNN42, CNN43, CNN44, CNN47, CNN48, CNN49, CNN50, CNN51, CNN52, CNN53, CNN55, CNN56, CNN57, CNN60, CNN61, CNN62, CNN63, CNN64, CNN65, CNN66, CNN67, CNN68, CNN70, CNN71, CNN72, CNN75, CNN76, CNN77, CNN78, CNN79, CNN80, CNN81, CNN82, CNN83, CNN84, CNN85, CNN86, CNN93, CNN94, CNN97, CNN98, CNN99, CNN100, CNN101, CNN103, CNN105, CNN107

*Shared genes are indicated only for one selected specie. See supplementary tables for the respective ORF numbers in each specie.

895 different ORFs, a set of sequences that might be called *the whole baculovirus gene content*. This high number of potential coding sequences contrasts with the range of gene content among the family members, which is between 90–181 genes (*Alphabaculovirus*: 118–169; *Betabaculovirus*: 116–181; *Gammabaculovirus*: 90–93; *Deltabaculovirus*: 109) as well as with the proportion of core genes which represents only 3%. This curious biological feature supports the hypothesis that highlights the great importance of structural mutations in the macroevolution of viruses with large DNA genomes. From this view, the set of genes shared by all members belonging to each baculovirus genus was compared to those corresponding to the *whole genus gene content* (Figure 3).

The analysis shows that Group I alphabaculoviruses and gammabaculoviruses have a lower diversity of gene content with respect to the rest of lineages. This information, coupled with the significant number of genome sequences obtained from Group I alphabaculoviruses, suggests that this lineage of viruses would constitute the newest clade in baculovirus evolution history [149]. This is based on the assumption that Group I alphabaculoviruses have had less time to incorporate new sequences from different sources (host genomes, other viral genomes, bacterial genomes, etc.) since the appearance of their common ancestor.

4. Baculovirus Core Gene Phylogeny

Traditional attempts to infer relationships between baculoviruses were performed by amino acid or nucleotide sequence analyses of single genes encoding proteins such as polyhedrin/granulin (the major component of OBs), the envelope fusion polypeptides known as F protein and GP64, or DNA polymerase protein, among many other examples [149–152].

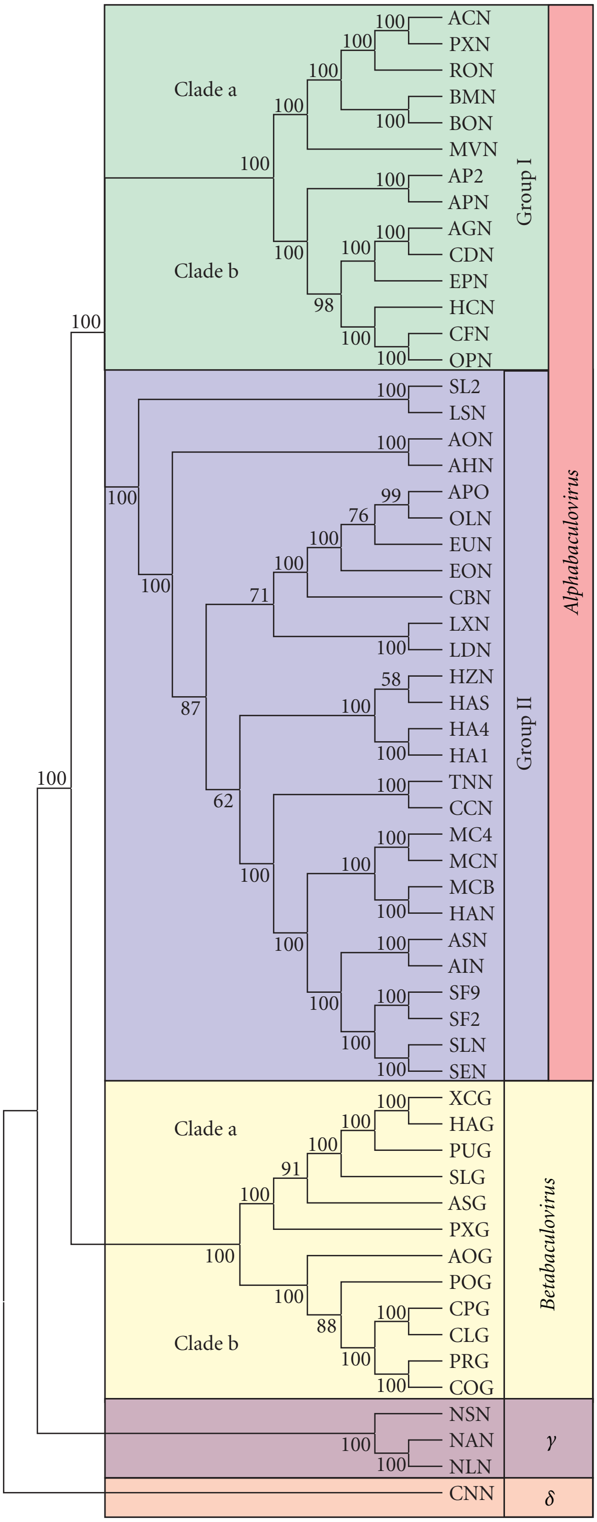


Figure 4: Baculovirus genome phylogeny. Cladogram based on amino acid sequence of core genes. The 31 identified core genes from Baculoviridae family were independently aligned using MEGA 4 [25] program with gap open penalty = 10, gap extension penalty = 1, and dayhoff matrix [26]. Then, a concatemer was generated and phylogeny inferred using the same software (UPGMA; bootstrap with 1000 replicates; gap/missing data = complete deletion; model = amino (dayhoff matrix); patterns among sites = same (homogeneous); rates among sites = different (gamma distributed); gamma parameter = 2.25). Baculoviruses are identified by the acronyms given in Table 1, and the accepted distribution in lineages and genera are also indicated. *Gammabaculovirus* and *Deltabaculovirus* are referenced by Greek letters. The proposed clades of Betabaculoviruses are shown in bold letters.

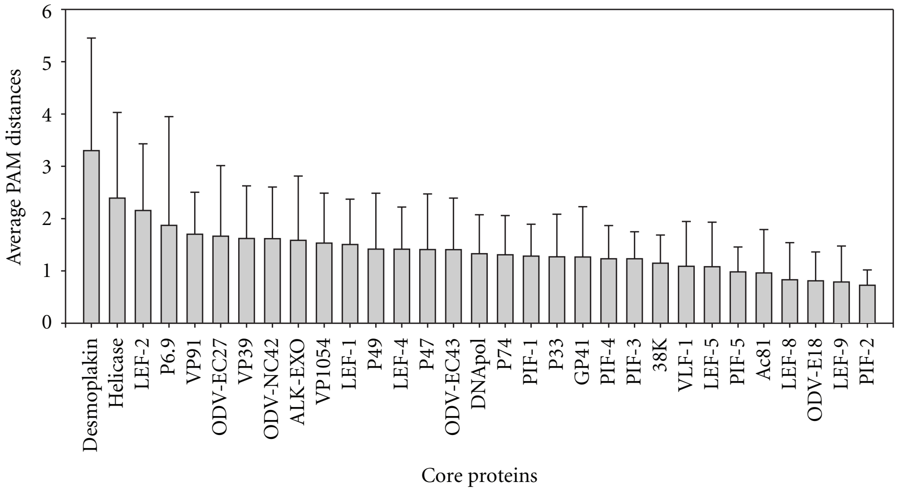


Figure 5: Baculovirus core gene variability. Histograms show the average PAM250 distances for each core gene with their corresponding standard deviations. These values were calculated using MEGA 4 program (UPGMA; bootstrap with 1000 replicates; gap/missing data = complete deletion; model = amino (dayhoff matrix); patterns among sites = same (homogeneous); rates among sites = different (gamma distributed; gamma parameter = 2.25)). PAM (point accepted mutation) matrices refers to the evolutionary distance between pairs of sequences. Given the weak similarity between several core proteins, PAM250 matrix was selected. The divergence considered in this matrix is 250 mutations per 100 amino acid sequence and was calculated to analyze more distantly related sequences. PAM250 is considered a good general matrix for protein similarity search.

Mostly, the evolutionary inferences were in agreement with much stronger subsequent studies based on sequence analyses derived from sets of genes with homologous sequences in all baculoviruses. Thus, these new approaches were based on the construction of common-protein-concatemers which were used to propose evolution patterns for baculoviruses [149].

Then, the fact that a viral family consists of members who share a common pattern of genes and functions and whose proliferation cycle continuously challenges the viral viability turns it essential to take into account their higher or lesser tolerance to the molecular changes. Molecular constraints regarding tolerance to changes in core genes are different from those of other genes. Therefore, core genes should be considered the most ancestral genes which may have diverged in higher or lesser degrees. According to this, a phylogenetic study was performed based on concatemers obtained from multiple alignments of the 31 proteins recognized in this work as core genes for the 57 available baculoviruses with sequenced genomes (Figure 4).

The obtained cladogram reproduces the current baculovirus classification based on 4 genera. Additionally, this approach consistently separates the alphabaculoviruses into two lineages: Group I and Group II. And the same can be observed when analyzing Group I, where the presence of two different clades can be clearly inferred (clade a and clade b). These groupings result in accordance with previous reports [20, 150]. In Group II alphabaculoviruses, a clear clustering may not be identified and would not allow to suggest a subdivision.

In contrast, in the *Betabaculovirus* genus, it is possible to propose their separation into two different clades: clade a (XnGV, HearGV, PsunGV, SpliGV, AgseGV, and PlxyGV), and clade b (AdorGV, PhopGV, CpGV, CrleGV, PiraGV, ChocGV).

Despite the evolutionary inference based on core genes, there was a remaining question: “is the tolerance to changes in all core genes the same?”. The answer could be reached by an individual core gene variability analysis for which studies of sequence distance for each baculovirus core gene were performed (Figure 5).

The resulting order of core genes shows that *pif-2* was the most conserved baculovirus ancestral sequence, whereas *desmoplakin* was the gene with evidence of greatest variability. This analysis reveals that genomes can be evolutionarily constrained in different ways depending on the proteins they encode.

The gain of access to new hosts might be an important force for gene evolution. During an infection process, the genome variants that appear with mutations introduced by errors in the replication/reparation machinery could be quickly incorporated into the virus population if the nucleotide changes offered a better biological performance when proteins were translated. The *DNA helicase* gene was considered as an important host range factor being, for this study, the second core sequence showing more variability [87]. However, other sequences like *pif-2* gene would not accumulate mutations because the protein encoded might lose vital functions not necessarily associated with the nature of the host.

5. Conclusions

Baculoviridae is a large family of viruses which infect and kill insect species from different orders. The valuable applications of these viruses in several fields of life sciences encourage their constant study with the goal of understanding the molecular mechanisms involved in the generation of progeny in the appropriate cells as well as the processes by which they evolve. The establishment of solid bases to recognize their phylogenetic relationships is necessary to facilitate the generation of new knowledge and the development of better methodologies.

In view of this, many researchers have proposed and used different bioinformatic methodologies to identify genes as well as related baculoviruses. Some of them were based on gene sequences [150], gene content [17], or genome rearrangements [152]. In this work, a combination of core gene sequence and gene content analyses were applied to reevaluate Baculoviridae classification. To our knowledge, the most important fact is that this report is the first work which identifies the whole baculovirus gene content and the shared genes that are unique in different genera and subgenera. All this information should be taken into account to group and classify new virus isolates and to propose molecular methodologies to diagnose baculoviruses based on proper gene targets according to gene variability and gene content.

Acknowledgments

This work was supported by research funds from *Agencia Nacional de Promoción Científica y Técnica* (ANPCyT) and *Universidad Nacional de Quilmes*. P. D. Ghiringhelli is member of the Research Career of CONICET (*Consejo Nacional de Investigaciones Científicas y Técnicas*), M. N. Belaich holds a postdoctoral fellowship of CONICET, S. A. B. Miele holds a fellowship of CONICET, and M J Garavaglia holds a fellowship of CIC-PBA (*Comisión de Investigaciones Científicas de la Provincia de Buenos Aires*). The authors acknowledge to Lic. Javier A. Iserte, Lic. Betina I. Stephan and Lic. Laura Esteban for their helping with the paper. S. A. B. Miele and M. Javier Garavaglia both contributed equally to this work.

References

[1] J. A. Jehle, G. W. Blissard, B. C. Bonning et al., “On the classification and nomenclature of baculoviruses: a proposal for revision,” *Archives of Virology*, vol. 151, no. 7, pp. 1257–1266, 2006.

[2] G. W. Blissard and G. F. Rohrmann, “Baculovirus diversity and molecular biology,” *Annual Review of Entomology*, vol. 35, no. 1, pp. 127–155, 1990.

[3] E. A. Kozlov, T. L. Levitina, and N. M. Gusak, “The primary structure of baculovirus inclusion body proteins. Evolution and structure-function aspects,” *Current Topics in Microbiology and Immunology*, vol. 131, pp. 135–164, 1986.

[4] G. F. Rohrmann, “Baculovirus structural proteins,” *Journal of General Virology*, vol. 73, no. 4, pp. 749–761, 1992.

[5] T. Ohkawa, J. O. Washburn, R. Sitapara, E. Sid, and L. E. Volkman, “Specific binding of *Autographa californica* M nucleopolyhedrovirus occlusion-derived virus to midgut cells of heliothis virescens larvae is mediated by products of pif genes Ac119 and Ac022 but not by Ac115,” *Journal of Virology*, vol. 79, no. 24, pp. 15258–15264, 2005.

[6] R. Jackes, E. Maromorosch, and K. Sherman, *Stability of Insect Viruses in the Environment. Viral Insecticides for Biological Control*, Academic Press, New York, NY, USA, 1985.

[7] G. Zhang, “Research, development and application of Heliothis viral pesticide in China,” *Resources and Environment in the Yangtze Valley*, vol. 3, pp. 1–6, 1994.

[8] R. D. Possee, “Baculoviruses as expression vectors,” *Current Opinion in Biotechnology*, vol. 8, no. 5, pp. 569–572, 1997.

[9] F. Moscardi, “Assessment of the application of baculoviruses for control of lepidoptera,” *Annual Review of Entomology*, vol. 44, pp. 257–289, 1999.

[10] T. A. Kost and J. P. Condreay, “Recombinant baculoviruses as expression vectors for insect and mammalian cells,” *Current Opinion in Biotechnology*, vol. 10, no. 5, pp. 428–433, 1999.

[11] A. B. Inceoglu, S. G. Kamita, A. C. Hinton et al., “Recombinant baculoviruses for insect control,” *Pest Management Science*, vol. 57, no. 10, pp. 981–987, 2001.

[12] T. A. Kost, J. P. Condreay, and D. L. Jarvis, “Baculovirus as versatile vectors for protein expression in insect and mammalian cells,” *Nature Biotechnology*, vol. 23, no. 5, pp. 567–575, 2005.

[13] M. D. Summers, “Milestones leading to the genetic engineering of baculoviruses as expression vector systems and viral pesticides,” *Advances in Virus Research*, vol. 68, pp. 3–73, 2006.

[14] A. B. Inceoglu, S. G. Kamita, and B. D. Hammock, “Genetically modified baculoviruses: a historical overview and future outlook,” *Advances in Virus Research*, vol. 68, pp. 323–360, 2006.

[15] X. Shi and D. L. Jarvis, “Protein N-glycosylation in the baculovirus-insect cell system,” *Current Drug Targets*, vol. 8, no. 10, pp. 1116–1125, 2007.

[16] J. P. Condreay and T. A. Kost, “Baculovirus expression vectors for insect and mammalian cells,” *Current Drug Targets*, vol. 8, no. 10, pp. 1126–1131, 2007.

[17] X. L. Sun and H. Y. Peng, “Recent advances in biological control of pest insect by using viruses in China,” *Virologica Sinica*, vol. 22, no. 2, pp. 158–162, 2007.

[18] E. A. Herniou, T. Luque, X. Chen et al., “Use of whole genome sequence data to infer baculovirus phylogeny,” *Journal of Virology*, vol. 75, no. 17, pp. 8117–8126, 2001.

[19] E. A. Herniou, J. A. Olszewski, J. S. Cory, and D. R. O’Reilly, “The genome sequence and evolution of baculoviruses,” *Annual Review of Entomology*, vol. 48, pp. 211–234, 2003.

[20] J. A. Jehle, M. Lange, H. Wang, Z. Hu, Y. Wang, and R. Hauschild, “Molecular identification and phylogenetic analysis of baculoviruses from Lepidoptera,” *Virology*, vol. 346, no. 1, pp. 180–193, 2006.

[21] M. M. van Oers and J. M. Vlak, “Baculovirus genomics,” *Current Drug Targets*, vol. 8, no. 10, pp. 1051–1068, 2007.

[22] G. F. Rohrman, *Baculovirus Molecular Biology*, National Library of Medicine (US), NCBI, Bethesda, Md, USA, 2008.

[23] T. Hayakawa, G. F. Rohrmann, and Y. Hashimoto, “Patterns of genome organization and content in lepidopteran baculoviruses,” *Virology*, vol. 278, no. 1, pp. 1–12, 2000.

[24] C. B. McCarthy and D. A. Theilmann, “AcMNPV ac143 (odv-e18) is essential for mediating budded virus production and is the 30th baculovirus core gene,” *Virology*, vol. 375, no. 1, pp. 277–291, 2008.

[25] K. Tamura, J. Dudley, M. Nei, and S. Kumar, “MEGA4: molecular evolutionary genetics analysis (MEGA) software version 4.0,” *Molecular Biology and Evolution*, vol. 24, no. 8, pp. 1596–1599, 2007.

[26] R. M. Schwartz and M. O. Dayhoff, "Matrices for detecting distant relationships," in *Atlas of Protein Sequences*, M. O. Dayhoff, Ed., pp. 353–358, National Biomedical Research Foundation, 1979.

[27] Q. Fan, S. Li, L. Wang et al., "The genome sequence of the multinucleocapsid nucleopolyhedrovirus of the Chinese oak silkworm Antheraea pernyi," *Virology*, vol. 366, no. 2, pp. 304–315, 2007.

[28] Z. M. Nie, Z. F. Zhang, D. Wang et al., "Complete sequence and organization of Antheraea pernyi nucleopolyhedrovirus, a dr-rich baculovirus," *BMC Genomics*, vol. 8, Article ID 248, 2007.

[29] J. V. de Castro Oliveira, J. L. C. Wolff, A. Garcia-Maruniak et al., "Genome of the most widely used viral biopesticide: anticarsia gemmatalis multiple nucleopolyhedrovirus," *Journal of General Virology*, vol. 87, no. 11, pp. 3233–3250, 2006.

[30] M. D. Ayres, S. C. Howard, J. Kuzio, M. Lopez-Ferber, and R. D. Possee, "The complete DNA sequence of *Autographa californica* nuclear polyhedrosis virus," *Virology*, vol. 202, no. 2, pp. 586–605, 1994.

[31] S. Gomi, K. Majima, and S. Maeda, "Sequence analysis of the genome of *Bombyx mori* nucleopolyhedrovirus," *Journal of General Virology*, vol. 80, no. 5, pp. 1323–1337, 1999.

[32] Y. P. Xu, Z. P. Ye, C. Y. Niu et al., "Comparative analysis of the genomes of Bombyx mandarina and *Bombyx mori* nucleopolyhedroviruses," *Journal of Microbiology*, vol. 48, no. 1, pp. 102–110, 2010.

[33] J. G. de Jong, H. A. M. Lauzon, C. Dominy et al., "Analysis of the Choristoneura fumiferana nucleopolyhedrovirus genome," *Journal of General Virology*, vol. 86, no. 4, pp. 929–943, 2005.

[34] H. A. M. Lauzon, P. B. Jamieson, P. J. Krell, and B. M. Arif, "Gene organization and sequencing of the Choristoneura fumiferana defective nucleopolyhedrovirus genome," *Journal of General Virology*, vol. 86, no. 4, pp. 945–961, 2005.

[35] O. Hyink, R. A. Dellow, M. J. Olsen et al., "Whole genome analysis of the *Epiphyas postvittana* nucleopolyhedrovirus," *Journal of General Virology*, vol. 83, no. 4, pp. 957–971, 2002.

[36] M. Ikeda, M. Shikata, N. Shirata, S. Chaeychomsri, and M. Kobayashi, "Gene organization and complete sequence of the *Hyphantria cunea* nucleopolyhedrovirus genome," *Journal of General Virology*, vol. 87, no. 9, pp. 2549–2562, 2006.

[37] Y. R. Chen, C. Y. Wu, S. T. Lee et al., "Genomic and host range studies of *Maruca vitrata* nucleopolyhedrovirus," *Journal of General Virology*, vol. 89, no. 9, pp. 2315–2330, 2008.

[38] C. H. Ahrens, R. L. Q. Russell, C. J. Funk, J. T. Evans, S. H. Harwood, and G. F. Rohrmann, "The sequence of the *Orgyia pseudotsugata* multinucleocapsid nuclear polyhedrosis virus genome," *Virology*, vol. 229, no. 2, pp. 381–399, 1997.

[39] R. L. Harrison and B. C. Bonning, "The nucleopolyhedroviruses of *Rachiplusia ou* and *Anagrapha falcifera* are isolates of the same virus," *Journal of General Virology*, vol. 80, no. 10, pp. 2793–2798, 1999.

[40] M. Nakai, C. Goto, W. Kang, M. Shikata, T. Luque, and Y. Kunimi, "Genome sequence and organization of a nucleopolyhedrovirus isolated from the smaller tea tortrix, *Adoxophyes honmai*," *Virology*, vol. 316, no. 1, pp. 171–183, 2003.

[41] S. Hilton and D. Winstanley, "Genomic sequence and biological characterization of a nucleopolyhedrovirus isolated from the summer fruit tortrix, *Adoxophyes orana*," *Journal of General Virology*, vol. 89, no. 11, pp. 2898–2908, 2008.

[42] A. K. Jakubowska, S. A. Peters, J. Ziemnicka, J. M. Vlak, and M. M. van Oers, "Genome sequence of an enhancin gene-rich nucleopolyhedovirus (NPV) from *Agrotis segetum*: collinearity with *Spodoptera exigua* multiple NPV," *Journal of General Virology*, vol. 87, no. 3, pp. 537–551, 2006.

[43] M. M. van Oers, M. H. C. Abma-Henkens, E. A. Herniou, J. C. W. de Groot, S. Peters, and J. M. Vlak, "Genome sequence of *Chrysodeixis chalcites* nucleopolyhedrovirus, a baculovirus with two DNA photolyase genes," *Journal of General Virology*, vol. 86, no. 7, pp. 2069–2080, 2005.

[44] S. Y. Zhu, J. P. Yi, W. D. Shen et al., "Genomic sequence, organization and characteristics of a new nucleopolyhedrovirus isolated from Clanis bilineata larva," *BMC Genomics*, vol. 10, Article ID 91, 9 pages, 2009.

[45] X. C. Ma, J. Y. Shang, Z. N. Yang, Y. Y. Bao, Q. Xiao, and C. X. Zhang, "Genome sequence and organization of a nucleopolyhedrovirus that infects the tea looper caterpillar, *Ectropis obliqua*," *Virology*, vol. 360, no. 1, pp. 235–246, 2007.

[46] X. D. Tang, Q. Xiao, X. C. Ma, Z. R. Zhu, and C. X. Zhang, "Morphology and genome of *Euproctis pseudoconspersa* nucleopolyhedrovirus," *Virus Genes*, vol. 38, no. 3, pp. 495–506, 2009.

[47] C. X. Zhang, X. C. Ma, and Z. J. Guo, "Comparison of the complete genome sequence between C1 and G4 isolates of the *Helicoverpa armigera* single nucleocapsid nucleopolyhedrovirus," *Virology*, vol. 333, no. 1, pp. 190–199, 2005.

[48] J. G. Ogembo, S. Chaeychomsri, K. Kamiya et al., "Cloning and comparative characterization of nucleopolyhedroviruses isolated from African bollworm, *Helicoverpa armigera*, (Lepidoptera: Noctudiae) in different geographic regions," *Journal of Insect Biotechnology and Sericology*, vol. 76, no. 1, pp. 39–49, 2007.

[49] X. Chen, W. F. J. Ijkel, C. Dominy et al., "Identification, sequence analysis and phylogeny of the lef-2 gene of *Helicoverpa armigera* single-nucleocapsid baculovirus," *Virus Research*, vol. 65, no. 1, pp. 21–32, 1999.

[50] H. Xiao and Y. Qi, "Genome sequence of *Leucania seperata* nucleopolyhedrovirus," *Virus Genes*, vol. 35, no. 3, pp. 845–856, 2007.

[51] J. Kuzio, M. N. Pearson, S. H. Harwood et al., "Sequence and analysis of the genome of a baculovirus pathogenic for *Lymantria dispar*," *Virology*, vol. 253, no. 1, pp. 17–34, 1999.

[52] Y. S. Nai, C. Y. Wu, T. C. Wang et al., "Genomic sequencing and analyses of *Lymantria xylina* multiple nucleopolyhedrovirus," *BMC Genomics*, vol. 11, no. 1, Article ID 116, 2010.

[53] S. Li, M. Erlandson, D. Moody, and C. Gillott, "A physical map of the *Mamestra configurata* nucleopolyhedrovirus genome and sequence analysis of the polyhedrin gene," *Journal of General Virology*, vol. 78, no. 1, pp. 265–271, 1997.

[54] L. Li, Q. Li, L. G. Willis, M. Erlandson, D. A. Theilmann, and C. Donly, "Complete comparative genomic analysis of two field isolates of *Mamestra configurata* nucleopolyhedrovirus-A," *Journal of General Virology*, vol. 86, no. 1, pp. 91–105, 2005.

[55] L. Li, C. Donly, Q. Li et al., "Identification and genomic analysis of a second species of nucleopolyhedrovirus isolated from *Mamestra configurata*," *Virology*, vol. 297, no. 2, pp. 226–244, 2002.

[56] R. L. Harrison, B. Puttler, and H. J. R. Popham, "Genomic sequence analysis of a fast-killing isolate of *Spodoptera frugiperda* multiple nucleopolyhedrovirus," *Journal of General Virology*, vol. 89, no. 3, pp. 775–790, 2008.

[57] J. L. C. Wolff, F. H. Valicente, R. Martins, J. V. Oliveira, and P. M. Zanotto, "Analysis of the genome of *Spodoptera frugiperda* nucleopolyhedrovirus (SfMNPV-19) and of the high genomic heterogeneity in group II nucleopolyhedroviruses," *Journal of General Virology*, vol. 89, no. 5, pp. 1202–1211, 2008.

[58] Y. Pang, J. Yu, L. Wang et al., "Sequence analysis of the *Spodoptera litura* multicapsid nucleopolyhedrovirus genome," *Virology*, vol. 287, no. 2, pp. 391–404, 2001.

[59] L. G. Willis, R. Siepp, T. M. Stewart, M. A. Erlandson, and D. A. Theilmann, "Sequence analysis of the complete genome of *Trichoplusia ni* single nucleopolyhedrovirus and the identification of a baculoviral photolyase gene," *Virology*, vol. 338, no. 2, pp. 209–226, 2005.

[60] S. Wormleaton, J. Kuzio, and D. Winstanley, "The complete sequence of the *Adoxophyes orana* granulovirus genome," *Virology*, vol. 311, no. 2, pp. 350–365, 2003.

[61] S. R. Escasa, H. A. M. Lauzon, A. C. Mathur, P. J. Krell, and B. M. Arif, "Sequence analysis of the *Choristoneura occidentalis* granulovirus genome," *Journal of General Virology*, vol. 87, no. 7, pp. 1917–1933, 2006.

[62] M. Lange and J. A. Jehle, "The genome of the *Cryptophlebia leucotreta* granulovirus," *Virology*, vol. 317, no. 2, pp. 220–236, 2003.

[63] T. Luque, R. Finch, N. Crook, D. R. O'Reilly, and D. Winstanley, "The complete sequence of the *Cydia pomonella* granulovirus genome," *Journal of General Virology*, vol. 82, no. 10, pp. 2531–2547, 2001.

[64] R. L. Harrison and H. J. R. Popham, "Genomic sequence analysis of a granulovirus isolated from the Old World bollworm, *Helicoverpa armigera*," *Virus Genes*, vol. 36, no. 3, pp. 565–581, 2008.

[65] A. Taha, A. Nour-el-Din, L. Croizier, M. López Ferber, and G. Croizier, "Comparative analysis of the granulin regions of the *Phthorimaea operculella* and *Spodoptera littoralis* granuloviruses," *Virus Genes*, vol. 21, no. 3, pp. 147–155, 2000.

[66] Y. Hashimoto, T. Hayakawa, Y. Ueno, T. Fujita, Y. Sano, and T. Matsumoto, "Sequence analysis of the *Plutella xylostella* granulovirus genome," *Virology*, vol. 275, no. 2, pp. 358–372, 2000.

[67] Y. Wang, J. Y. Choi, J. Y. Roh, S. D. Woo, B. R. Jin, and Y. H. Je, "Molecular and phylogenetic characterization of *Spodoptera litura* granulovirus," *Journal of Microbiology*, vol. 46, no. 6, pp. 704–708, 2008.

[68] T. Hayakawa, R. Ko, K. Okano, S. I. Seong, C. Goto, and S. Maeda, "Sequence analysis of the *Xestia c-nigrum* granulovirus genome," *Virology*, vol. 262, no. 2, pp. 277–297, 1999.

[69] S. P. Duffy, A. M. Young, B. Morin, C. J. Lucarotti, B. F. Koop, and D. B. Levin, "Sequence analysis and organization of the Neodiprion abietis nucleopolyhedrovirus genome," *Journal of Virology*, vol. 80, no. 14, pp. 6952–6963, 2006.

[70] H. A. M. Lauzon, C. J. Lucarotti, P. J. Krell, Q. Feng, A. Retnakaran, and B. M. Arif, "Sequence and organization of the *Neodiprion lecontei* nucleopolyhedrovirus genome," *Journal of Virology*, vol. 78, no. 13, pp. 7023–7035, 2004.

[71] H. A. M. Lauzon, A. Garcia-Maruniak, P. M. de A. Zanotto et al., "Genomic comparison of Neodiprion sertifer and *Neodiprion lecontei* nucleopolyhedroviruses and identification of potential hymenopteran baculovirus-specific open reading frames," *Journal of General Virology*, vol. 87, no. 6, pp. 1477–1489, 2006.

[72] A. Garcia-Maruniak, J. E. Maruniak, P. M. A. Zanotto et al., "Sequence analysis of the genome of the Neodiprion sertifer nucleopolyhedrovirus," *Journal of Virology*, vol. 78, no. 13, pp. 7036–7051, 2004.

[73] C. L. Afonso, E. R. Tulman, Z. Lu et al., "Genome sequence of a baculovirus pathogenic for *Culex nigripalpus*," *Journal of Virology*, vol. 75, no. 22, pp. 11157–11165, 2001.

[74] J. T. Evans, D. J. Leisy, and G. F. Rohrmann, "Characterization of the interaction between the baculovirus replication factors LEF-1 and LEF-2," *Journal of Virology*, vol. 71, no. 4, pp. 3114–3119, 1997.

[75] A. L. Vanarsdall, K. Okano, and G. F. Rohrmann, "Characterization of the replication of a baculovirus mutant lacking the DNA polymerase gene," *Virology*, vol. 331, no. 1, pp. 175–180, 2005.

[76] J. Huang and D. B. Levin, "Expression, purification and characterization of the *Spodoptera littoralis* nucleopolyhedrovirus (SpliNPV) DNA polymerase and interaction with the SpliNPV non-hr origin of DNA replication," *Journal of General Virology*, vol. 82, no. 7, pp. 1767–1776, 2001.

[77] V. V. McDougal and L. A. Guarino, "*Autographa californica* nuclear polyhedrosis virus DNA polymerase: measurements of processivity and strand displacement," *Journal of Virology*, vol. 73, no. 6, pp. 4908–4918, 1999.

[78] X. Hang and L. A. Guarino, "Purification of *Autographa californica* nucleopolyhedrovirus DNA polymerase from infected insect cells," *Journal of General Virology*, vol. 80, no. 9, pp. 2519–2526, 1999.

[79] J. G. M. Heldens, Y. Liu, D. Zuidema, R. W. Goldbach, and J. M. Vlak, "Characterization of a putative *Spodoptera exigua* multicapsid nucleopolyhedrovirus helicase gene," *Journal of General Virology*, vol. 78, no. 12, pp. 3101–3114, 1997.

[80] S. Maeda, S. G. Kamita, and A. Kondo, "Host range expansion of *Autographa californica* nuclear polyhedrosis virus (NPV) following recombination of a 0.6-kilobase-pair DNA fragment originating from *Bombyx mori* NPV," *Journal of Virology*, vol. 67, no. 10, pp. 6234–6238, 1993.

[81] E. Ito, D. Sahri, R. Knippers, and E. B. Carstens, "Baculovirus proteins IE-1, LEF-3, and P143 interact with DNA in vivo: a formaldehyde cross-linking study," *Virology*, vol. 329, no. 2, pp. 337–347, 2004.

[82] V. V. Mcdougal and L. A. Guarino, "The *Autographa californica* nuclear polyhedrosis virus p143 gene encodes a DNA helicase," *Journal of Virology*, vol. 74, no. 11, pp. 5273–5279, 2000.

[83] D. K. Bideshi and B. A. Federici, "DNA-independent ATPase activity of the *Trichoplusia ni* granulovirus DNA helicase," *Journal of General Virology*, vol. 81, no. 6, pp. 1601–1604, 2000.

[84] GE. Liu and E. B. Carstens, "Site-directed mutagenesis of the AcMNPV p 143 gene: effects on baculovirus DNA replication," *Virology*, vol. 253, no. 1, pp. 125–136, 1999.

[85] D. K. Bideshi and B. A. Federici, "The *Trichoplusia ni* granulovirus helicase in unable to support replication of *Autographa californica* multicapsid nucleopolyhedrovirus in cells and larvae of T. ni," *Journal of General Virology*, vol. 81, no. 6, pp. 1593–1599, 2000.

[86] J. T. Evans, G. S. Rosenblatt, D. J. Leisy, and G. F. Rohrmann, "Characterization of the interaction between the baculovirus ssDNA-binding protein (LEF 3) and putative helicase (P143)," *Journal of General Virology*, vol. 80, no. 2, pp. 493–500, 1999.

[87] O. Argaud, L. Croizier, M. López-Ferber, and G. Croizier, "Two key mutations in the host-range specificity domain of the p143 gene of *Autographa californica* nucleopolyhedrovirus are required to kill *Bombyx mori* larvae," *Journal of General Virology*, vol. 79, no. 4, pp. 931–935, 1998.

[88] S. G. Kamita and S. Maeda, "Abortive infection of the baculovirus *Autographa californica* nuclear polyhedrosis virus in Sf-9 cells after mutation of the putative DNA helicase gene," *Journal of Virology*, vol. 70, no. 9, pp. 6244–6250, 1996.

[89] G. Croizier, L. Croizier, O. Argaud, and D. Poudevigne, "Extension of *Autographa californica* nuclear polyhedrosis virus host range by interspecific replacement of a short DNA sequence in the p143 helicase gene," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 91, no. 1, pp. 48–52, 1994.

[90] G. Liu and E. B. Carstens, "Site-directed mutagenesis of the AcMNPV p 143 gene: effects on baculovirus DNA replication," *Virology*, vol. 253, no. 1, pp. 125–136, 1999.

[91] L. A. Guarino, B. Xu, J. Jin, and W. Dong, "A virus-encoded RNA polymerase purified from baculovirus-infected cells," *Journal of Virology*, vol. 72, no. 10, pp. 7985–7991, 1998.

[92] L. A. Guarino, J. Jin, and W. Dong, "Guanylyltransferase activity of the LEF-4 subunit of baculovirus RNA polymerase," *Journal of Virology*, vol. 72, no. 12, pp. 10003–10010, 1998.

[93] C. H. Gross and S. Shuman, "RNA 5'-triphosphatase, nucleoside triphosphatase, and guanylyltransferase activities of baculovirus LEF-4 protein," *Journal of Virology*, vol. 72, no. 12, pp. 10020–10028, 1998.

[94] J. Jin, W. Dong, and L. A. Guarino, "The LEF-4 subunit of baculovirus RNA polymerase has RNA 5'- triphosphatase and ATPase activities," *Journal of Virology*, vol. 72, no. 12, pp. 10011–10019, 1998.

[95] C. H. Gross and S. Shuman, "Characterization of a baculovirus-encoded RNA 5'-triphosphatase," *Journal of Virology*, vol. 72, no. 9, pp. 7057–7063, 1998.

[96] J. S. Titterington, T. K. Nun, and A. L. Passarelli, "Functional dissection of the baculovirus late expression factor-8 gene: sequence requirements for late gene promoter activation," *Journal of General Virology*, vol. 84, no. 7, pp. 1817–1826, 2003.

[97] C. Iorio, J. E. Vialard, S. McCracken, M. Lagacé, and C. D. Richardson, "The late expression factors 8 and 9 and possibly the phosphoprotein p78/83 of *Autographa californica* multicapsid nucleopolyhedrovirus are components of the virus-induced RNA polymerase," *Intervirology*, vol. 41, no. 1, pp. 35–46, 1998.

[98] J. R. McLachlin and L. K. Miller, "Identification and characterization of vlf-1, a baculovirus gene involved in very late gene expression," *Journal of Virology*, vol. 68, no. 12, pp. 7746–7756, 1994.

[99] A. Lu and L. K. Miller, "The roles of eighteen baculovirus late expression factor genes in transcription and DNA replication," *Journal of Virology*, vol. 69, no. 2, pp. 975–982, 1995.

[100] J. W. Todd, A. L. Passarelli, A. Lu, and L. K. Miller, "Factors regulating baculovirus late and very late gene expression in transient-expression assays," *Journal of Virology*, vol. 70, no. 4, pp. 2307–2317, 1996.

[101] A. L. Passarelli and L. K. Miller, "Identification of genes encoding late expression factors located between 56.0 and 65.4 map units of the *Autographa californica* nuclear polyhedrosis virus genome," *Virology*, vol. 197, no. 2, pp. 704–714, 1993.

[102] M. E. Wilson and L. K. Miller, "Changes in the nucleoprotein complexes of a baculovirus DNA during infection," *Virology*, vol. 151, no. 2, pp. 315–328, 1986.

[103] M. E. Wilson, T. H. Mainprize, P. D. Friesen, and L. K. Miller, "Location, transcription, and sequence of a baculovirus gene encoding a small arginine-rich polypeptide," *Journal of Virology*, vol. 61, no. 3, pp. 661–666, 1987.

[104] M. Wang, E. Tuladhar, S. Shen et al., "Specificity of baculovirus P6.9 basic DNA-binding proteins and critical role of the C terminus in virion formation," *Journal of Virology*, vol. 84, no. 17, pp. 8821–8828, 2010.

[105] S. M. Thiem and L. K. Miller, "Identification, sequence, and transcriptional mapping of the major capsid protein gene of the baculovirus *Autographa californica* nuclear polyhedrosis virus," *Journal of Virology*, vol. 63, no. 5, pp. 2008–2018, 1989.

[106] M. N. Pearson, R. L. Q. Russell, G. F. Rohrmann, and G. S. Beaudreau, "p39, a major baculovirus structural protein: immunocytochemicalcharacterization and genetic location," *Virology*, vol. 167, no. 2, pp. 407–413, 1988.

[107] G. W. Blissard, R. L. Quant-Russell, G. F. Rohrmann, and G. S. Beaudreau, "Nucleotide sequence, transcriptional mapping, and temporal expression of the gene encoding p39, a major structural protein of the multicapsid nuclear polyhedrosis virus of *Orgyia pseudotsugata*," *Virology*, vol. 168, no. 2, pp. 354–362, 1989.

[108] S. Lu, G. Ge, and Y. Qi, "Ha-VP39 binding to actin and the influence of F-actin on assembly of progeny virions," *Archives of Virology*, vol. 149, no. 11, pp. 2187–2198, 2004.

[109] J. R. McLachlin and L. K. Miller, "Identification and characterization of vlf-1, a baculovirus gene involved in very late gene expression," *Journal of Virology*, vol. 68, no. 12, pp. 7746–7756, 1994.

[110] S. Yang and L. K. Miller, "Control of baculovirus polyhedrin gene expression by very late factor 1," *Virology*, vol. 248, no. 1, pp. 131–138, 1998.

[111] S. Yang and L. K. Miller, "Expression and mutational analysis of the baculovirus very late factor 1 (vlf-1) gene," *Virology*, vol. 245, no. 1, pp. 99–109, 1998.

[112] V. S. Mikhailov and G. F. Rohrmann, "Binding of the baculovirus very late expression factor 1 (VLF-1) to different DNA structures," *BMC Molecular Biology*, vol. 3, Article ID 14, 2002.

[113] A. L. Vanarsdall, K. Okano, and G. F. Rohrmann, "Characterization of the role of very late expression factor 1 in baculovirus capsid structure and DNA processing," *Journal of Virology*, vol. 80, no. 4, pp. 1724–1733, 2006.

[114] V. S. Mikhailov, K. Okano, and G. F. Rohrmann, "Baculovirus alkaline nuclease possesses a 5′ → 3′ exonuclease activity and associates with the DNA-binding protein LEF-3," *Journal of Virology*, vol. 77, no. 4, pp. 2436–2444, 2003.

[115] V. S. Mikhailov, K. Okano, and G. F. Rohrmann, "Specificity of the endonuclease activity of the baculovirus alkaline nuclease for single-stranded DNA," *Journal of Biological Chemistry*, vol. 279, no. 15, pp. 14734–14745, 2004.

[116] K. Okano, A. L. Vanarsdall, and G. F. Rohrmann, "Characterization of a baculovirus lacking the alkaline nuclease gene," *Journal of Virology*, vol. 78, no. 19, pp. 10650–10656, 2004.

[117] J. Olszewski and L. K. Miller, "Identification and characterization of a baculovirus structural protein, VP1054, required for nucleocapsid formation," *Journal of Virology*, vol. 71, no. 7, pp. 5040–5050, 1997.

[118] R. L. Q. Russell and G. F. Rohrmann, "Characterization of P91, a protein associated with virions of an *Orgyia pseudotsugata* baculovirus," *Virology*, vol. 233, no. 1, pp. 210–223, 1997.

[119] M. Whitford and P. Faulkner, "A structural polypeptide of the baculovirus *Autographa californica* nuclear polyhedrosis virus contains O-linked N-acetylglucosamine," *Journal of Virology*, vol. 66, no. 6, pp. 3324–3329, 1992.

[120] J. Olszewski and L. K. Miller, "A role for baculovirus GP41 in budded virus production," *Virology*, vol. 233, no. 2, pp. 292–301, 1997.

[121] W. Wu, T. Lin, L. Pan et al., "*Autographa californica* multiple nucleopolyhedrovirus nucleocapsid assembly is interrupted upon deletion of the 38K gene," *Journal of Virology*, vol. 80, no. 23, pp. 11475–11485, 2006.

[122] W. Wu, H. Liang, J. Kan et al., "*Autographa californica* multiple nucleopolyhedrovirus 38K is a novel nucleocapsid protein that interacts with VP1054, VP39, VP80, and itself," *Journal of Virology*, vol. 82, no. 24, pp. 12356–12364, 2008.

[123] C. M. Long, G. F. Rohrmann, and G. F. Merrill, "The conserved baculovirus protein p33 (Ac92) is a flavin adenine dinucleotide-linked sulfhydryl oxidase," *Virology*, vol. 388, no. 2, pp. 231–235, 2009.

[124] W. Wu and A. L. Passarelli, "*Autographa californica* multiple nucleopolyhedrovirus Ac92 (ORF92, P33) is required for budded virus production and multiply enveloped occlusion-derived virus formation," *Journal of Virology*, vol. 84, no. 23, pp. 12351–12361, 2010.

[125] Y. Nie, M. Fang, and D. A. Theilmann, "*Autographa californica* multiple nucleopolyhedrovirus core gene ac92 (p33) is required for efficient budded virus production," *Virology*, vol. 409, no. 1, pp. 38–45, 2011.

[126] M. Fang, H. Wang, H. Wang et al., "Open reading frame 94 of *Helicoverpa armigera* single nucleocapsid nucleopolyhedrovirus encodes a novel conserved occlusion-derived virion protein, ODV-EC43," *Journal of General Virology*, vol. 84, no. 11, pp. 3021–3027, 2003.

[127] F. Deng, R. Wang, M. Fang et al., "Proteomics analysis of *Helicoverpa armigera* single nucleocapsid nucleopolyhedrovirus identified two new occlusion-derived virus-associated proteins, HA44 and HA100," *Journal of Virology*, vol. 81, no. 17, pp. 9377–9385, 2007.

[128] KE. Peng, M. Wu, F. Deng et al., "Identification of protein-protein interactions of the occlusion-derived virus-associated proteins of *Helicoverpa armigera* nucleopolyhedrovirus," *Journal of General Virology*, vol. 91, no. 3, pp. 659–670, 2010.

[129] A. L. Vanarsdall, M. N. Pearson, and G. F. Rohrmann, "Characterization of baculovirus constructs lacking either the Ac 101, Ac 142, or the Ac 144 open reading frame," *Virology*, vol. 367, no. 1, pp. 187–195, 2007.

[130] G. Li, J. Wang, R. Deng, and X. Wang, "Characterization of AcMNPV with a deletion of ac68 gene," *Virus Genes*, vol. 37, no. 1, pp. 119–127, 2008.

[131] C. B. McCarthy and D. A. Theilmann, "AcMNPV ac143 (odv-e18) is essential for mediating budded virus production and is the 30th baculovirus core gene," *Virology*, vol. 375, no. 1, pp. 277–291, 2008.

[132] J. Ke, J. Wang, R. Deng, and X. Wang, "*Autographa californica* multiple nucleopolyhedrovirus ac66 is required for the efficient egress of nucleocapsids from the nucleus, general synthesis of preoccluded virions and occlusion body formation," *Virology*, vol. 374, no. 2, pp. 421–431, 2008.

[133] M. Belyavskyi, S. C. Braunagel, and M. D. Summers, "The structural protein ODV-EC27 of *Autographa californica* nucleopolyhedrovirus is a multifunctional viral cyclin," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 95, no. 19, pp. 11205–11210, 1998.

[134] S. C. Braunagel, H. He, P. Ramamurthy, and M. D. Summers, "Transcription, translation, and cellular localization of three *Autographa californica* nuclear polyhedrosis virus structural proteins: ODV-E18, ODV-E35, and ODV-EC27," *Virology*, vol. 222, no. 1, pp. 100–114, 1996.

[135] H. Q. Chen, KE. P. Chen, Q. Yao, Z. J. Guo, and L. L. Wang, "Characterization of a late gene, ORF67 from *Bombyx mori* nucleopolyhedrovirus," *FEBS Letters*, vol. 581, no. 30, pp. 5836–5842, 2007.

[136] O. Simón, S. Gutiérrez, T. Williams, P. Caballero, and M. López-Ferber, "Nucleotide sequence and transcriptional analysis of the pif gene of *Spodoptera frugiperda* nucleopolyhedrovirus (SfMNPV)," *Virus Research*, vol. 108, no. 1-2, pp. 213–220, 2005.

[137] G. P. Pijlman, A. J. P. Pruijssers, and J. M. Vlak, "Identification of pif-2, a third conserved baculovirus gene required for per os infection of insects," *Journal of General Virology*, vol. 84, no. 8, pp. 2041–2049, 2003.

[138] P. Faulkner, J. Kuzio, G. V. Williams, and J. A. Wilson, "Analysis of p74, a PDV envelope protein of *Autographa californica* nucleopolyhedrovirus required for occlusion body infectivity in vivo," *Journal of General Virology*, vol. 78, no. 12, pp. 3091–3100, 1997.

[139] W. Zhou, L. Yao, H. Xu, F. Yan, and Y. Qi, "The function of envelope protein p74 from *Autographa californica* multiple nucleopolyhedrovirus in primary infection to host," *Virus Genes*, vol. 30, no. 2, pp. 139–150, 2005.

[140] E. J. Haas-Stapleton, J. O. Washburn, and L. E. Volkman, "P74 mediates specific binding of *Autographa californica* M nucleopolyhedrovirus occlusion-derived virus to primary cellular targets in the midgut epithelia of Heliothis virescens larvae," *Journal of Virology*, vol. 78, no. 13, pp. 6786–6791, 2004.

[141] L. Yao, W. Zhou, H. Xu, Y. Zheng, and Y. Qi, "The *Heliothis armigera* single nucleocapsid nucleopolyhedrovirus envelope protein P74 is required for infection of the host midgut," *Virus Research*, vol. 104, no. 2, pp. 111–121, 2004.

[142] S. C. Braunagel, W. K. Russell, G. Rosas-Acosta, D. H. Russell, and M. D. Summers, "Determination of the protein composition of the occlusion-derived virus of *Autographa californica* nucleopolyhedrovirus," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, no. 17, pp. 9797–9802, 2003.

[143] I. Kikhno, S. Gutiérrez, L. Croizier, G. Crozier, and M. López Ferber, "Characterization of pif, a gene required for the per os infectivity of *Spodoptera littoralis* nucleopolyhedrovirus," *Journal of General Virology*, vol. 83, no. 12, pp. 3013–3022, 2002.

[144] T. Ohkawa, J. O. Washburn, R. Sitapara, E. Sid, and L. E. Volkman, "Specific binding of *Autographa californica* M nucleopolyhedrovirus occlusion-derived virus to midgut cells of heliothis virescens larvae is mediated by products of pif genes Ac119 and Ac022 but not by Ac115," *Journal of Virology*, vol. 79, no. 24, pp. 15258–15264, 2005.

[145] M. Fang, Y. Nie, S. Harris, M. A. Erlandson, and D. A. Theilmann, "*Autographa californica* multiple nucleopolyhedrovirus core gene ac96 encodes a per os infectivity factor (pif-4)," *Journal of Virology*, vol. 83, no. 23, pp. 12569–12578, 2009.

[146] S. C. Braunagel, D. M. Elton, H. Ma, and M. D. Summers, "Identification and analysis of an *Autographa californica* nuclear polyhedrosis virus structural protein of the occlusion-derived virus envelope: ODV-E56," *Virology*, vol. 217, no. 1, pp. 97–110, 1996.

[147] W. O. Sparks, R. L. Harrison, and B. C. Bonning, "*Autographa californica* multiple nucleopolyhedrovirus ODV-E56 is a per os infectivity factor, but is not essential for binding and fusion of occlusion-derived virus to the host midgut," *Virology*, vol. 409, no. 1, pp. 69–76, 2011.

[148] D. P. A. Cohen, M. Marek, B. G. Davies, J. M. Vlak, and M. M. van Oers, "Encyclopedia of *Autographa californica* nucleopolyhedrovirus genes," *Virologica Sinica*, vol. 24, no. 5, pp. 359–414, 2009.

[149] Y. Jiang, F. Deng, S. Rayner, H. Wang, and Z. Hu, "Evidence of a major role of GP64 in group I alphabaculovirus evolution," *Virus Research*, vol. 142, no. 1-2, pp. 85–91, 2009.

[150] E. A. Herniou and J. A. Jehle, "Baculovirus phylogeny and evolution," *Current Drug Targets*, vol. 8, no. 10, pp. 1043–1050, 2007.

[151] P. M. de Andrade Zanotto and D. C. Krakauer, "Complete genome viral phylogenies suggests the concerted evolution of regulatory cores and accessory satellites," *PLoS ONE*, vol. 3, no. 10, Article ID e3500, 2008.

[152] D. Goodman, N. Ollikainen, and C. Sholley, "Baculovirus phylogeny based on genome rearrangements," in *Proceedings of the International Conference on Comparative Genomics*, vol. 4751 of *Lecture Notes in Computer Science*, pp. 69–82, 2007.