Gene 598 (2017) 63–70

Contents lists available at ScienceDirect

Gene

journal homepage: www.elsevier.com/locate/gene

Research paper

# Analysis of gene gain and loss in the evolution of predatory bacteria

Nan Li [a,b,*], Kai Wang [c], Henry N Williams [b], Jun Sun [a], Changling Ding [a], Xiaoyun Leng [a], Ke Dong [d]

[a] College of Marine Science and Environmental Sciences, Tianjin University of Science and Technology, Tianjin 300457, China
[b] School of the Environment, Florida A&M University, Tallahassee, FL, USA
[c] Institute of Biochemistry and Cell Biology, Shanghai Institute of Biological Science, Chinese Academy of Science, Shanghai, China
[d] Department of Biological Sciences, Seoul National University, Seoul 151-742, South Korea

ARTICLE INFO

*Article history:*
Received 12 May 2016
Received in revised form 11 October 2016
Accepted 27 October 2016
Available online 5 November 2016

*Keywords:*
Gene gain and loss
BALOs
*Bacteriovorax*
*Bdellovibrio*
*Micavibrio*
*Halobacteriovorax*
Positive selection
Comparative genomics

ABSTRACT

Predatory bacteria are ubiquitously distributed in nature in including in aquatic environments, sewage, intestinal tracts of animals and humans, rhizophere and, soils. However, our understanding of their evolutionary history is limited. Results of recent studies have shown that acquiring novel genes is a major force driving bacterial evolution. Therefore, to gain a better understanding of the impact of gene gain and loss in the evolution of bacterial predators, this study employed comparative genomic approaches to identify core-set gene families and species-specific gene families, and model gene gain and loss events among 11 genomes that represented diverse lineages. In total, 1977 gene families were classified. Of these 509 (pattern 11111111111) were present all of the 11 species. Among the non-core set gene families, 52 were present only in saltwater bacteria predators and had no ortholog in the other genomes. Similarly 109 and 44 were present only in the genomes of *Micavibrio* spp. and *Bdellovibrio* spp., respectively. In this study, the gain loss mapping engine GLOOME was selected to analyze and estimate the expectations and probabilities of both gain and loss events in the predatory bacteria. In total, 354 gene families were involved in significant gene gain events, and 407 gene families were classified into gene loss events with high supported value. Moreover, 18 families from the core set gene family were identified as putative genes under positive selection. The results of this study suggest that acquisition of particular genes that encode functional proteins in metabolism and cellular processes and signaling, especially ABC systems, may help bacterial predators adapt to surrounding environmental changes and present different predation strategies for survival in their habitats.



## 1. Introduction

Bacterial evolution can be driven by frequent gene gain and loss occurrences within gene families. Although identical genus/species share a common core gene set, individuals within the genus/species may vary in the subset of genes they possess (Dobrindt and Hacker, 2001; Bubic et al., 2004; Marri et al., 2006). This subset of genes may hold the key for the ability of a bacterium to survive in a particular habitat. Such adaptation-specific variation can be brought about by gene gain resulting from lateral gene transfer (LGT) (Boucher and Doolittle, 2000; Gogarten et al., 2002; Ochman et al., 2005; Ragan and Beiko, 2009), or gene loss (Cole et al., 2001; Carlton et al., 2007), or modification of existing genes (Sokurenko et al., 1999; Feldgarden et al., 2003). Genes that arise from LGT may encode new functions some of which may result in adaptation to a changing environment (Hao and Golding, 2004). Gain and loss of genetic material has long been recognized by evolutionary biologists to be an important process augmenting site-specific mutations in the evolution of species of bacteria (Achtman and Wagner, 2008). For example, it was recently reported that the 248 putative secreted effectors in *Melanopsichium pennsylvanicum* smut genomes might represent a core set required for pathogenicity. On the other hand, the 92 secreted effectors common in grass-parasitic smuts but which show no ortholog in *Me. pennsylvanicum* may represent a set of effectors required for colonization of their grass hosts (Sharma et al., 2014). Another example of how gene gain and loss events may have played a role in evolution is a case with *Prochlorococcus*. The 1273 genes shared among 12 strains of *Prochlorococcus* were reported to encode the requisites of a functional cell capable of producing living matter by the process of photosynthesis (Kettler et al., 2007).

Predatory bacteria are ubiquitously distributed in nature including in all aquatic systems, sewage, soils, rhizophere, and the intestinal tracts of various animals. The *Bdellovibrio* and like organisms (BALOs) are perhaps the best characterized predatory bacteria. They are obligate predators that can only survive by preying on other gram-negative bacteria. BALOs consist of a freshwater group, *Bdellovibrio, Peredibacter,*

*Abbreviations:* LGT, lateral gene transfer; BALOs, *Bdellovibrio* and like organisms; IMG, Integrated Microbial Genomes; COGs, orthologous groups of proteins; MSA, multiple sequence alignment; RSD, Reciprocal Smallest Distance.

* Corresponding author at: Florida A& M University, 1515 S Martin Luther King, Jr Blvd, Tallahassee, FL 32307, USA.
*E-mail addresses:* nan.li@famu.edu, nli0417@gmail.com (N. Li).

http://dx.doi.org/10.1016/j.gene.2016.10.039
0378-1119/

*and Bacteriovorax* and a saltwater group, *Halobacteriovorax* (Koval et al., 2015),known previously as *Bacteriovorax*. In addition to these taxa, there are other known predators, which are yet to be characterized at the phylogenetic level (Fig. 1).

Both ancient and recent LGT has been previously reported in predatory bacteria Order, *Bdellovibrionales* (Gophna et al., 2006; Pan et al., 2011; Crossman et al., 2013). However, questions about which of the lateral transferred genes within the bacterial predators are significant, and which are unnecessary are still not well known. Also due to the limitation of available predatory bacterial genomes at the time, the evolutionary trace among distinct groups of predatory bacteria has not been well addressed. To more fully assess diversification and adaptation in the evolution of predatory bacteria, available sequences of 11 genomes of diverse lineages from *Bdellovibrionales* and other bacterial predators from the Integrated Microbial Genomes (IMG) system (Markowitz et al., 2008, 2012) were compared. Our objective was to reconstruct the history of gene acquisition and gene loss for these predators and specifically, those functions associated with the core and flexible genomes. This analysis revealed not only what differentiates bacteria predators, but also provides essential biological contexts of how adaptation occurs in predatory events in fresh water and salt water organisms.

## 2. Materials and methods

### 2.1. Sequences used

The study involved 11 predatory bacteria genome sequences: *Halobacteriovorax* spp. BSW11 (PRJNA210325), *H.* spp. DB6_IX (PRJNA210327), *H. marinus* SJ (PRJNA50431), *H.* spp. Seq25_V (PRJNA210326), *H.* spp. BAL6_X (PRJNA210328), *Micavibrio aeruginosavorus* ARL-13 (PRJNA49751), *M.* spp. EPB (PRJNA163341), *Bdellovibrio exovorus* JSS (PRJNA163339), *B. bacteriovorus* HD100 (PRJNA9637), *B. bacteriovorus Tiberius* (PRJNA70801), *Saprospira grandis* Lewin (PRJNA67697). All the sequences were downloaded from the Integrated Microbial Genomes (IMG) system (Markowitz et al., 2008, 2012).

### 2.2. Phylogenetic analysis

The multiple alignments of 16S rRNA from the 11 bacterial predator genomes were conducted using the MUSCLE v3.8.31(Edgar, 2004) with default parameters. Phylogenetic trees were constructed using NJ methods of the MEGA package (Version 6.0) (Tamura et al., 2013), and the reliability of each branch was determined by 1000 bootstrap replications. The concatenated DNA sequences, obtained by joining the individual single-copy genes represented in the 11 genomes, were used to reconstruct the phylogeny of the various taxa of bacteria predators. The sequences were aligned using Probalign (Roshan and Livesay, 2006). The maximum likelihood analysis of the concatenated DNA sequences, were performed by PhyML program (Guindon et al., 2009, 2010). This tree was used for further gene gain and loss analysis. To circumvent the confounding effects of duplication, which may occur during evolution (Zhang and He, 2005; Gu and Su, 2007), paralogs of gene families were not included in the phylogeny construction.

### 2.3. Identification of core-set gene families

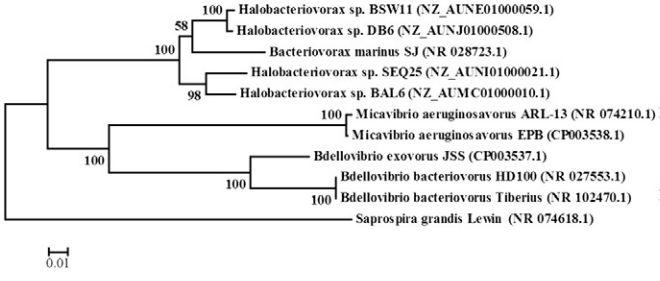
The sequences were first classified into gene families. The potential homologs were identified by sequence similarities, and all paralogs in each genome were clustered as a single gene family. The protein sequences from all genomes were compared using BlastP (Altschul et al., 1990; Mount, 2007), with an E-value cutoff set at $10^{-10}$ and an additional criterion of match length set at 85% of the query sequence. In order to further analyze the gene family, CD-HIT program (Li and Godzik, 2006; Fu et al., 2012) was used to check for clustering the gene families with 0.85 as a cutoff value. The sequences present in all the strains were recognized as the core-set gene family. The gene families belonged to a single species and were not found in any of the other bacteria predator genomes and were, therefore, considered species-specific. The sequences were annotated using clusters of orthologous groups of proteins (COGs) database (Tatusov et al., 2000) by BlastP (Altschul et al., 1990; Mount, 2007) with an E-value cutoff of $10^{-10}$.

### 2.4. Gene gain and gene loss

To examine the genes lost or gained in the genomes of bacteria predators during evolution, a gap-free multiple sequence alignment (MSA), in which rows correspond to species and columns correspond to binary characters ('1' denotes presence and '0' absence), was generated by the methods described in previous studies (Hao and Golding, 2004; Kettler et al., 2007; Cohen et al., 2010). Orthologs were considered to be absent in one genome if they were present in all other ten genomes, but not predicted in the genome being considered. A gene was considered to be species-specific if it was found solely in the species under consideration and no orthologs were detected in the other genomes. Gene presence and absence were further tested using the gain loss mapping engine GLOOME (Cohen et al., 2010). The expectations and probabilities of both gain and loss events were estimated using stochastic mapping.

### 2.5. Positive selection analysis

Protein-coding genes were extracted from each genome sequence, and orthologous protein clusters were delimited as described above. Orthologous gene content information was used to delimit the core genome of bacteria predators (509 orthologs). The alignments for each of the one-to-one core genes (509 orthologs) were generated with one outgroup sequence in the alignment, which, whenever possible, was *Saprospira grandis Lewin* (PRJNA67697). First, the sequences were




| | Habitats | Cluster* | Predatory strategy | Bioproject Accession | Genome Size (Mb) | Gene Count | GC% | CDS% |
|---|---|---|---|---|---|---|---|---|
| Halobacteriovorax sp. BSW11 (NZ_AUNE01000059.1) | Marine | IV | Periplasmic | PRJNA210325 | 3.6 | 3490 | 37 | 99.05 |
| Halobacteriovorax sp. DB6 (NZ_AUNJ01000508.1) | Marine | IX | Periplasmic | PRJNA210327 | 3 | 3217 | 38 | 99.22 |
| Bacteriovorax marinus SJ (NR 028723.1) | Marine | III | Periplasmic | PRJNA50431 | 3.4 | 3288 | 37 | 98.84 |
| Halobacteriovorax sp. SEQ25 (NZ_AUNI01000021.1) | Marine | V | Periplasmic | PRJNA210326 | 3.4 | 3331 | 36 | 98.83 |
| Halobacteriovorax sp. BAL6 (NZ_AUMC01000010.1) | Marine | X | Periplasmic | PRJNA210328 | 3.2 | 3100 | 36 | 98.87 |
| Micavibrio aeruginosavorus ARL-13 (NR 074210.1) | Fresh water/Soil | --- | Epibiotic | PRJNA49751 | 2.4 | 2476 | 55 | 98.26 |
| Micavibrio aeruginosavorus EPB (CP003538.1) | Fresh water | --- | Epibiotic | PRJNA163341 | 2.4 | 2412 | 55 | 98.22 |
| Bdellovibrio exovorus JSS (CP003537.1) | Sewage | --- | Periplasmic | PRJNA163339 | 2.6 | 2656 | 42 | 98.61 |
| Bdellovibrio bacteriovorus HD100 (NR 027553.1) | Fresh water | --- | Periplasmic | PRJNA9637 | 3.7 | 3636 | 51 | 98.65 |
| Bdellovibrio bacteriovorus Tiberius (NR 102470.1) | Fresh water/Soil | --- | Periplasmic | PRJNA70801 | 3.9 | 3779 | 50 | 98.92 |
| Saprospira grandis Lewin (NR 074618.1) | Marine | --- | Ixotrophy | PRJNA67697 | 4.4 | 4324 | 46 | 98.54 |

**Fig. 1.** The phylogenetic tree of the bacterial predators. The phylogenetic tree was constructed based on their 16S rRNA using the Neighbor-Joining method. The reliability of the tree was evaluated with 1000 replicates of bootstrapping test and only high bootstrap value scores (>50%) were indicated on the branches. In addition, each strain is followed by its living habitat, total number of genes, as well as absolute and relative number of other information. *Cluster were identified by previous study (Pineiro et al., 2007). 16S rRNA sequences of strains BSW 11, DB6, SEQ25 and BAL6 were extracted from their genomic sequences according to the annotation.