Pangenome Evolution in the Marine Bacterium Alteromonas

Mario López-Pérez and Francisco Rodriguez-Valera*

Evolutionary Genomics Group, Departamento de Producción Vegetal y Microbiología, Universidad Miguel Hernández, Alicante, Spain

*Corresponding author: E-mail: frvalera@umh.es.

Accepted: April 24, 2016

Data deposition: The genome sequences have been deposited in GenBank under the following BioProject accession number PRJNA307085.

Abstract

We have examined a collection of the free-living marine bacterium Alteromonas genomes with cores diverging in average nucleotide identities ranging from 99.98% to 73.35%, i.e., from microbes that can be considered members of a natural clone (like in a clinical epidemiological outbreak) to borderline genus level. The genomes were largely syntenic allowing a precise delimitation of the core and flexible regions in each. The core was 1.4 Mb (ca. 30% of the typical strain genome size). Recombination rates along the core were high among strains belonging to the same species (37.7–83.7% of all nucleotide polymorphisms) but they decreased sharply between species (18.9–5.1%). Regarding the flexible genome, its main expansion occurred within the boundaries of the species, i.e., strains of the same species already have a large and diverse flexible genome. Flexible regions occupy mostly fixed genomic locations. Four large genomic islands are involved in the synthesis of strain-specific glycosydic receptors that we have called glycotypes. These genomic regions are exchanged by homologous recombination within and between species and there is evidence for their import from distant taxonomic units (other genera within the family). In addition, several hotspots for integration of gene cassettes by illegitimate recombination are distributed throughout the genome. They code for features that give each clone specific properties to interact with their ecological niche and must flow fast throughout the whole genus as they are found, with nearly identical sequences, in different species. Models for the generation of this genomic diversity involving phage predation are discussed.

Key words: Alteromonas, pangenome, genomic islands, recombination, intraspecies diversity, phages.

Introduction

The Neodarwinian evolutionary paradigm (Maynard 1989) has little value for the analysis of the evolution of prokaryotes (Koonin and Wolf 2012). Prokaryotes have two-tiered genomes formed by a core of genes, that have homologs in all the strains within the species, and a flexible set characteristic of each strain that have no homologs in most or none of the others (Welch et al. 2002; Tettelin et al. 2008). The flexible pool is often associated to horizontal gene transfer (HGT) and site-specific recombination catalyzed by mobile genetic elements, whereas the core tends to be identified with vertical transmission and homologous recombination (Wiederbeck and Cohan 2011; Kuenne et al. 2013; Polz et al. 2013). However, it is better to define these two genomic compartments as evolutionary trends. Thus, the flexible pool provides diversity in the population, e.g., at the level of surface features or transporters that interact with the biotic and abiotic components of the environment, whereas the core provides a stable metabolic and genomic backbone for the population (Rodriguez-Valera and Ussery 2012). A significant part of the flexible genome is concentrated in genomic islands of > 10 kb. We use the term flexible genomic island (fGI) to refer the clusters of flexible genes that occupy equivalent genomic locations in different strains within a taxonomic unit (normally the species, although here we have used the term at the genus level) (Gonzaga et al. 2012; Chan et al. 2015) to distinguish them from the unique islands that are found in either one or the other strain. These FGIs can be classified into two major categories (López-Pérez et al. 2013; Chan et al. 2015). Additive FGIs that are site directed recombination hotspots for the integration of gene cassettes that vary in their numbers and nature (Kuenne et al. 2013). Their functional role is variable depending on the combination of cassettes they carry. Replacement FGIs are gene clusters related to polysaccharides or glycoproteins present in the outer layers of the cell. They are coded by completely different gene clusters in different strains, with no evidence of similarity (except for the overall functional annotation) but code for equivalent exposed structures. Replacement FGIs are instrumental for the environmental interactions of the cell that, in addition, acquires through...
them a surface identity. For example, they allow specific host recognition by phages (Rodriguez-Valera et al. 2009; Avrani et al. 2012). In pathogenic bacteria, in which this diversity has been extensively explored and described (Reeves 1993; Milkman 1997) they are recognized by the vertebrate immune system and generate a characteristic serotype. However, for free-living microbes this terminology seems out of place. In this work, we call the specific glycosydic receptors coded by these sets of genes “glycotype.” They are exchanged by homologous recombination facilitated by the conserved neighboring genes (Milkman et al. 2003; Jeong et al. 2009; Dingle et al. 2012; López-Pérez, Martin-Cuadrado, et al. 2014).

For some years, we have studied the marine bacterium Alteromonas with the focus on understanding its genomic make-up and dynamics (Gonzaga et al. 2012; López-Pérez et al. 2012, 2013; López-Pérez, Gonzaga, et al. 2014). This microbe is a heterotrophic marine gammaproteobacterium that is found in the open ocean and takes advantage of the sporadic inputs of organic matter that appear in this relatively impoverished environment. It is a typical “bloomer” that grows very fast and can be easily retrieved in pure culture. Although initially we focused on a single species (A. macleodii) this group of strains turned out to be made up of two species: Alteromonas macleodii and A. mediterranea, that display <84% average nucleotide identity (ANI) between them (Ivanova et al. 2015). We compared two genomes of A. mediterranea isolated from the same sample and compared them to a fosmid metagenomic library also from the same Mediterranean seawater sample. This study (Gonzaga et al. 2012) showed that the two isolates, members of the same population, had similar cores (over 98.51% ANI) but differed broadly in gene content of their fGIs. The presence of fosmids with overlapping sections indicated that there were at least five coexisting clones in the population. When we added to the comparison isolates from a different Mediterranean location and obtained at a different time (5 years apart) we still detected members of the same clones (López-Pérez et al. 2013). These isolates had evidence of very close common ancestry [<100 single nucleotide polymorphism (SNPs) throughout the core genome], and had identical glycoype determinants, but differed slightly in their additive fGIs. These highly similar genomes belonging to independent isolates were defined as members of the same clonal frame (CF). They are the smallest unit of differentiation of prokaryotes (often referred to as strains, biotypes, or serotypes). This term is also equivalent with “epidemic clone” in clinical microbiology (Cho et al. 2010; Mutreja et al. 2011; Cui et al. 2015).

With the addition of other species, A. macleodii (López-Pérez et al. 2012) and A. australica (López-Pérez, Gonzaga, et al. 2014), the genomes revealed remarkable conservation of synteny, which has facilitated the study of the physical arrangement of the core and flexible regions in the genomes of representatives of this genus (López-Pérez, Gonzaga, et al. 2014). In Alteromonas, four replacement fGIs were identified (López-Pérez et al. 2012, 2013; López-Pérez, Gonzaga, et al. 2014). They are associated to the synthesis of structural polysaccharides in the outer surface of the cell, specifically, flagellum glycosylation, capsular exopolysaccharide (EPS) and O-chain polysaccharide (two separate clusters) of the cell-wall lipopolysaccharide. Each different CF has a different glycoype with different versions for each of the four glycoypey determinants (López-Pérez et al. 2013). In addition, several additive fGIs were identified coding for multiple functions from H2 oxidation to polysaccharide metabolism.

We considered that the availability of this gradient of genomic diversity provided an opportunity to understand its origin and evolution within the genus Alteromonas. The results of this analysis could be extrapolated to other free-living aquatic bacteria. Therefore, we have analyzed Alteromonas genomes through comparative genomics, starting from the closest strains (same CF), and then adding gradually more distant taxonomic entities (strains belonging to different CFs within the same species and strains belonging to different species). Specifically, for this work, we have sequenced eight new genomes; we have also collected strains from strain collections and genomes available in public databases. All the genomes analyzed here are fully sequenced and assembled. They have been compared to dissect the mechanistic processes of their evolution using diverse bioinformatic tools. Finally, we describe a model of diversity fixation in populations of closely related bacteria by bacteriophage predation.

Materials and Methods
Sample Collection, Sequencing, Assembly and Annotation
Details of isolation and origin of the Alteromonas strains used are provided in supplementary table S1, Supplementary Material online. Strains R10SW13, LMG 21861 and LMG 21856 have been obtained from the culture collection of the DSMZ (https://www.dsmz.de/home.html). DNA was extracted by phenol-chloroform as described in (Neumann et al. 1992) and checked for quality on a 1% agarose gel. The quantity was measured using Quant-iT PicoGreen dsDNA Reagent (Invitrogen). Sequencing was performed using Illumina Hiseq 2000 (100-bp paired-end read) (BGI Tech Solutions, Hong Kong). The generated reads were trimmed and assembled de novo using the IDBA assembler (Peng et al. 2012). We used a combination of Geneious Pro 5.0.1 (with default parameters) using previously Alteromonas assembled genomes as a reference (Gonzaga et al. 2012; López-Pérez et al. 2012, 2013) and oligonucleotides designed from the sequence of the ends of assembled contigs to obtain one single-closed contig. The genomes were annotated using the NCBI PGAAP annotation pipeline.
The predicted protein sequences were compared using BLASTP to the NCBI nr protein database (e-value $10^{-5}$). ORFs <100 bp and without significant homology to other proteins were not considered. Reciprocal BLASTN and tblastx searches between the genomes were carried out, leading to the identification of regions of similarity, insertions and rearrangements. The ANI between strains was calculated using JSpecies software package v1.2.1 using default parameters (Richter and Rosselló-Móra 2009). To allow the interactive visualization of genomic fragment comparisons Artemis v.12 and Artemis Comparison Tool ACTv.9 were used (Carver et al. 2005, 2012). Additional annotation was carried out using the RAST server (Aziz et al. 2008). Additional local BLAST searches against the latest NCBI nr database were performed whenever necessary. Clustering algorithm CD-HIT (Fu et al. 2012) was performed with cut-offs of > 65% minimal alignment coverage and >85% sequence identity for each sequence pair in order to assign the sequences to orthologous clusters.

Phylogenetic Analysis

Whole-genome alignments for all 25 strains (plus Pseudalteromonas atlantica T6c that was used as an outgroup) were constructed using the progressiveMauve algorithm (Darling et al. 2010) of the Mauve software v2.3.1. The resulting alignments were subsequently used in ClonalFrame software v1.2 (Didelot et al. 2010). ClonalFrame was a Bayesian inference method which jointly reconstructs the clonal relationships between the isolates in a sample. The clonal genealogy inferred from the generated alignment of 1.38 Mbp of the core genome by ClonalFrame is shown in fig. 1. StripSubsetLCBs script, also including in the Mauve software v2.3.1, was used to extract all the regions of at least 1000 bp found in all the genomes and maximum-likelihood trees using RAxML (version 7.2.6) (Stamatakis 2014) were generated using these regions in order to confirm the low level of recombination detected among different species.

Evolutionary Rate and SNPs Analysis

ClustalW was used to align the sequence of orthologous proteins and program pal2nal program (Suyama et al. 2006) to obtain the multiple codon alignment from the corresponding aligned protein sequences in order to evaluate the type and rate of nucleotide substitutions. For each sequence pair, dN/dS ratios were calculated based on the codon alignments using SNAP (Korber 2000). A low ratio ($dN/dS < 1$) indicates purifying selection, whereas a high ratio ($dN/dS > 1$) is a clear signal of diversifying selection. For each pair of strains, pairwise alignments were performed using progressiveMauve algorithm (Darling et al. 2010). The resulting locally collinear blocks of at least 500 bp were extracted from the output of the program and concatenated to form the core genome alignment.

DnaSP 5.10.01 (Librado and Rozas 2009) was used to obtain the total number of SNPs between the genomes and SNPs in windows of 1 kb. Indels and SNPs between small regions of the genome such as genomic islands were identified using nucmer program in the MUMmer3+ package (Kurtz et al. 2004).

Results

Genomes Studied

We have analyzed genome sequences of Alteromonas species isolated over a 50-year span and from all around the world (supplementary table S1, Supplementary Material online), although we do not believe there is any biogeographic signal in this microbe (López-Pérez et al. 2013; López-Pérez, Gonzaga, et al. 2014). In total, we have studied here 25 fully assembled genomes classified into four clear-cut genospecies. We define genospecies as clusters of strains with genomes having >95% ANI (Konstantinidis and Tiedje 2005). Specifically, we have analyzed representatives of A. mediterranea (formerly A. macleodii deep ecotype) (11 strains), A. macleodii (7 strains), “A. stellipolaris” (5 strains) and A. australica (2 strains). For A. mediterranea, we have analyzed several strains that belong to the same CF. They provide a gradient of variation that goes from a few SNPs (ANI 99.98%), for members of the same CF, to borderline values for strains belonging to the same genus (ANI 73.35%) (supplementary fig. S1, Supplementary Material online). Eight strain genomes (U10, UM8, D7, Mac1, Mac2, LMG 21861, LMG 21856, and R105SW13) are reported here for the first time. The Ionian isolate UM8 was formerly considered identical to UM7 (López-Pérez et al. 2013) but has been now shown to be different and is therefore described for the first time. Two strains from polar latitudes A. stellipolaris (LMG 21861 and LMG 21856) (Van Trappen et al. 2004) and A. addita R105W13 (described as different species based on 16S rRNA) (Ivanova et al. 2005), together with two new Mediterranean isolates described here for the first time (Alteromonas sp. Mac1 and Mac2), belong to a single genospecies that we will call “A. stellipolaris” (the oldest denomination).

Overall Conservation of Synteny

In order to analyze the phylogenetic relationships of all the strains, we have clustered them using a whole-genome phylogeny from a concatenation of shared genes (ca. 1.38 Mbp) (fig. 1). The genomes analyzed have relatively similar sizes (ranging from 4.34 to 4.94 Mb, plasmids included). The strains cluster into four genospecies (ANI > 95%) (fig. 1 and supplementary fig. S1, Supplementary Material online). Synteny was quite well conserved throughout the genus and even the positions of the main features of the flexible genome were conserved, as has been observed before in other bacteria within this range of ANI variation (Kuenne et al. 2013; López-Pérez,
Replacement fGIs coding for the glyctype are highlighted in fig. 1 and the different versions of the replacement fGIs coding for them assigned a number.

The main variation in synteny was detected in all the “A. stellipolaris” isolates between positions 1–2.5 Mb (affecting ca. 30% of the genome) (fig. 1 and supplementary fig. S2, Supplementary Material online). These strains had a large rearrangement located in the second quadrant between the second and third ribosomal operon. Within this large rearrangement, we identified a large inversion (ca. 0.5 Mb).
869 kb) that included both the flagellum glycosylation and EPS fGIs (supplementary fig. S2, Supplementary Material online). Although the position of the flagellum glycosylation fGI appeared inverted, the position relative to the origin of replication of the EPS fGI (very close to the center of the inversion) remains similar to that of the other species. Using DE1 as a reference, the rearrangement breakpoint was located at a 4-kb intergenic region at the boundary of the flagellum glycosylation island and in the other side in a region enriched in IS elements. There are two additional small inversions of ca. 74 and 120 kb, respectively (supplementary fig. S2, Supplementary Material online).

The conservation of syntheny and the availability of more than one strain for each species allowed us to assess the patterns of genomic variation within and between species and eventually try to understand the evolutionary mechanisms behind. A summary of the divergence parameters of some of the pairwise genome comparisons is shown in table 1.

**Evolution of the Core Genome**

**Intraclonalframe**

We have compared the genomes by incremental steps of divergence starting by isolates that belong to a single CF. Alteromonas mediterranea CF1 and CF2 were described previously (López-Pérez et al. 2013). Now we have included more strains and analyzed in more detail the differences among them. Whole genome alignment among the five members of *A. mediterranea* CF1 (AltDE1, UM7, UM8, UM4b and U10) revealed a total of 151 homologous regions representing a core genome of 4.2 Mb (97% of the average genome size) and a total of 146 variable sites that include both SNPs and insertions or deletions.

Following the approach used for *Escherichia coli* by (Dixit et al. 2015), we have calculated different parameters based on SNP density to gauge the relative evolutionary divergence among these strains (summarized in table 1). Aligned regions between genomes were divided into 1 kb segments and divergence was estimated as the percentage of SNPs ($\Delta S$). The clonal fraction (nonrecombined) of each pair was identified as the segment fraction containing 0–3 SNPs ($\Delta R$), while the remaining, at least in part, to recombination ($\Delta R^*$. The clonal fraction/$\Delta M$; LCA, age of last common ancestor determined using the evolutionary rate of *Legionella pneumophila* (Sanchez-Buso et al. 2014); n.d., not determined.

Values that deviate from the analysis described in Dixit et al. (2015) are in parentheses.
UM4b, the most different within CF1 (ANI 99.92/Coverage 95.66% to DE1), appears to have suffered genome reduction, losing the plasmid pAMDE1-300 and ca. 200 kb of the chromosome (including the lysogenic phage). However, only 43 of the SNPs detected in this strain could be ascribed to mutation, indicating a similar divergence time with the other strains. The ratios of SNPs attributable to synonymous and nonsynonymous replacements (table 1) were similar for all of the CF1 strains, ca. 25% were located in intergenic regions. These figures support a recent divergence that has not allowed time for selection to act. To translate mutational SNP values into divergence time, we have used the evolutionary rate of the nonrecombining core genome of Legionella pneumophila (Sanchez-Buso et al. 2014), another aquatic gammaproteobacterium, as reference. We calculated that ca. 80 years (table 1) have elapsed since the last common ancestor of CF1 strains.

Similar figures were found for CF2, the two strains U4 and U8 core genome revealed 109 mutational SNPs in coding regions, of which 49 (45%) were nonsynonymous (table 1). These two strains should be separated by ca. 196 years since their last common ancestor (table 1). A careful examination of the variable sites within members of the same CF indicated that most could be the product of intragenomic duplication, showing enrichment in proteins neighboring the replacement fGIs (supplementary fig. S3, Supplementary Material online), typically involving tandem repeats of different lengths, gene duplications and the movement of transposable elements.

Intragenic segment transitions in the nonrecombining core genome of Legionella pneumophila (Sanchez-Buso et al. 2014), another aquatic gammaproteobacterium, as reference. We calculated that ca. 80 years (table 1) have elapsed since the last common ancestor of CF1 strains.

Intraspecies

When the comparison was shifted to strains belonging to different CFs within the same species, clonal fractions decreased to 0.624–0.040 as ∆S varied from nearly 0 to 1.66% (table 1). Besides, the total number of SNPs over the core genome varied from a few hundred to a range from 13,461 (LMG 21861-Mac1) to 66,301 (ATCC 27126-AD45). We aligned the core of four genome pairs, one for each Alteromonas genospecies, covering the whole range of intraspecies ANI values (supplementary fig. S3, Supplementary Material online). We plotted the SNP density (in 1 kb segments) (Dixit et al. 2015) for each genospecies pair (supplementary fig. S3, Supplementary Material online). They showed a similar distribution to those found when comparing different strains of E. coli (Dixit et al. 2015), having a maximum at the average frequency of mutational changes per kb, one for pairs within “A. stellipolaris” (LMG 21861-R10SW13) and A. mediterranea (DE1-U4); four for A. australica (H17-DE170) and nine for A. macleodii (ATCC 27126-AD45), respectively. Recently, it has been proposed for E. coli (Dixit et al. 2015) that clonal segments disappear at levels of divergence at or above 1 SNP/kb. Accepting this premise, in the pairs ATCC 27126-AD45 (ANI 96.75) and H17-DE170 (ANI 98.63), all clonal vertically-inherited parts are lost (supplementary fig. S3, Supplementary Material online). Interestingly this happens at about the same overall divergence (ANI 98.8) as in E. coli (Dixit et al. 2015). The most divergent intraspecies pairs (<98% ANI) also followed a Poisson distribution with peak values > 1 SNP/kb (supplementary fig. S3, Supplementary Material online). Although after the previously mentioned model (Dixit et al. 2015), these genomes are likely to be completely covered by a mosaic of recombined segments (∆M = 0%; ∆R = 100%), we posit that fragments with small amounts of SNPs (located to the left of the peak) could be due to mutation and the rest (to the right of the peak) to recombination. If this assumption is correct, ∆R varied from 93% (LMG21861-R10SW13, ANI 98.8%) to 37% (ATCC 27126-AD45, ANI 96.75) (table 1). These values are similar to those found for intraspecies recombination in other Gammaproteobacteria (Vos and Didelot 2009).

In order to analyze the impact of homologous recombination in the evolution of the core genome, we estimated the rtm parameter that measures the relative weight of recombination to mutation in sequence divergence (table 1). Results showed that for all intragenomic pairs, mostly intraspecies comparisons rtm was >1, i.e., recombination introduced more substitutions than mutation, with the highest value observed between U4 and U8 members of the CF2 (table 1). However, in the two most divergent intraspecies pairs (H17-DE170 and ATCC 27126-AD45), rtm was significantly <1 (more mutation than recombination). These data suggest that recombination rates decrease proportionally to ANI values.

The last common ancestor of the different CFs within each species calculated from ∆M values, as above, was between 4,000 and 74,000 years old (table 1). Annotation of the segments with the highest numbers of SNPs attributable to recombination, showed enrichment in proteins neighboring the replacement fGIs (supplementary fig. S3, Supplementary Material online), known to be frequent targets of recombination (López-Pérez et al. 2013). The same kind of genes has been found to be highly recombinogenic in E. coli (Dixit et al. 2015). Other genes overrepresented in the high recombination tails coded for proteins that are exposed in the outermost cell structures like TonB receptors, porins or pillins. They might also be involved in environmental or phage interaction.

To gauge the selection pressure acting on these strain pairs, we estimated the ratio of nonsynonymous to synonymous substitution rates dN/dS. The values for all the intraspecies comparisons (table 1) were very similar, with a mean of 0.69 ± 0.04, indicating weak stabilizing selection, similar values were obtained for different serovars of the bacterial pathogen Salmonella enterica Agona (Zhou et al. 2013), Typhi (Holt et al. 2008) and Typhimurium (Hawkey et al. 2013) and other pathogenic bacteria. We also examined the dN/dS variation throughout the genome. A whole genome alignment was constructed with the five A. mediterranea genomes belonging to different CFs partitioning the alignment into 10 kb windows. This process generated 348 fragments containing 3.4 Mb of core aligned nucleotides. Among these
fragments, 81 (23%) had higher dN than dS. Similar values were obtained for the strains within A. macleodii. Supplementary table S2, Supplementary Material online, lists the genes with the highest dN/dS for A. mediterranea strains. Among the 20 genes associated with highest dN/dS values eight were transporters of different kinds, including TonB receptors. This might reflect a divergence in nutrient ranges utilized by different strains within the same species.

Interspecies

The cores of strains belonging to different genospecies were much more divergent, A. macleodii and A. mediterranea with an ANI ca. 84% had ΔS values ca. 9.2% while A. australica and “A. stellipolaris” (ANI ca. 73%) 13% (table 1). Despite the 10-fold increase in ΔS compared with the intraspecies values, the dN/dS ratio was similar, 0.65 for the first group and ca. 0.72 for the more distant genospecies. Thus, it appears that purifying selection is already saturated at the intraspecies level. The variation of the SNP density plots with increasing ANI distance is shown in fig. 2, from CFs to interspecies, using as reference the genome of A. mediterranea DE1. It is apparent that recombination is much rarer between species (ΔR ca. 5% compared with ca. 60% within species). This reduction has been described in other Gammaproteobacteria such as Vibrio (Urbanczyk et al. 2015). Given that for the most divergent pairs the model predicting recombination from SNP density is not applicable (Dixit et al. 2015), to assess recombination rates at the interspecies level, we generated ML trees of different alignable genomic regions and compared them to the consensus tree generated by the core genome of all the strains (supplementary fig. S4, Supplementary Material online). The results illustrate that recombination events often break the topology within the species but not between species. All these analysis support the decline in homologous recombination between increasingly divergent genomes that, as has been suggested before, might lead to prokaryotic speciation (Shapiro et al. 2012).

Flexible Genome

Flexible Genome Expansion

The pairwise variation in shared genes among all the Alteromonas strains (all vs. all) is shown in fig. 3. The shared gene clusters (core) and differential (flexible) for all couples of strains have reciprocal values, as could be expected given that the genome size is relatively constant. In addition, as expected, both values were proportional to the ANI values, with the number of common clusters increasing, and the numbers of not shared clusters decreasing, at higher ANI values (fig. 3A). It is remarkable that most of the expansion of the flexible genome takes place within the species range of variation, i.e., some pairs belonging to the same species have as many different genes as pairs of strains belonging to different species. Furthermore, even when comparing pairs belonging to the same CF the numbers of nonshared genes was quite significant (in the most extreme values close to 500). Figure 3B plot shows the incremental change of core and flexible with each of the genomes added. Typically each new strain within the species added 300 new genes while the species change added between 500 and 700. Together all the strains (considering only one member of each CF) were shown to possess a pangenome of 9,623 genes (1,795 core and 7,828 flexible) (fig. 3B).

We have analyzed the locations where synteny was lost across the genomes due to gene acquisition and loss for the two species of Alteromonas with a significant number of strains each (A. macleodii and A. mediterranea). Using as reference strain DE1 of A. mediterranea, we identified only 25 synteny breaks when its genome was compared with the five strains belonging to the same CF (fig. 4, upper panel). By contrast, when compared with the other five CFs within A. mediterranea, the number of synteny breaks increased to 507 (fig. 4, middle panel). When the comparison was extended to all the other strains in the genus (including the other species) the number of synteny breaks increased only to 584 (fig. 4, lower panel). It seems that the sites for synteny breaks were largely exhausted within the species boundaries.

Thus, the real qualitative jump in the flexible genome diversity in Alteromonas happens at the intraspecies level. We found a similar number of synteny breaks (494) for the seven strains in A. macleodii (supplementary fig. S5, Supplementary Material online). These breaks contain 1,783 and 1,496 genes.
in *A. mediterranea* and *A. macleodii*, respectively. Synteny breaks involving segments of <10 kb were ca. 93% of the total number of synteny breaks in both species, but they contained only around half of the flexible genes (49% and 59%, respectively). The other part of the flexible genome was present in segments >10 kb (largely fGIs), 30 in *A. mediterranea* and 34 in *A. macleodii*. The biological role of the larger synteny breaks and their relative location appeared conserved throughout the genus (fig. 5).

**Exchange of Alteromonas Glycotype Gene Clusters**

Figure 1 shows the location and different versions (identified by a number) of the four replacement fGIs (glycotype gene clusters) described in *Alteromonas* (Gonzaga et al. 2012; López-Pérez et al. 2012, 2013). Each different CF had a totally different set of these islands. The exceptions are highlighted in red in fig. 1. The case of “*A. stellipolaris*” can be considered special because the flagellum glycosylation island and the O-chain main gene cluster were highly similar although the five
strains belong to three different CFs. However, the EPS and the O-chain cluster 2 were different for each CF as is normally the case, their relationship at ANI level was also atypical when compared with the other species with intermediate values between the intraclonal frame and intraspecies level.

It is known that replacement fGIs are exchanged by homologous recombination (Jeong et al. 2009; López-Pérez et al. 2013; López-Pérez, Martin-Cuadrado, et al. 2014). We already had detected exchange between different CFs (a nearly identical version of the flagellum glycosylation island was found in two different strains within *A. mediterranea* (López-Pérez et al. 2013). In the collection of strains presented here, more examples of exchange between different CFs, and even species, could be detected (fig. 1). It is important to emphasize, however, that recombination of glycoype determinants in *Alteromonas* happens at a much lower rate than the exchange of ecologically relevant characters found in additive islands (López-Pérez et al. 2013; López-Pérez, Martin-Cuadrado, et al. 2014). Firstly, in CF1 of *A. mediterranea* the gene clusters were identical, i.e., likely to remain vertically transmitted, although the strains have diverged for a large number of generations (see above), and were pretty divergent at the level of other components of the flexible genome. Secondly, the frequency of SNPs when comparing the putatively exchanged gene clusters indicates that thousands of generations elapse between these homologous recombination events of exchange. For example, in *A. macleodii*, the strain pairs 673-AD006 and AD45-AD037 share the same
flagellum cluster but with 116 SNPs over 18.4 kb and 926 SNPs over 28.8 kb, respectively. These values are indicative of old exchanges.

Although the most frequently exchanged fGI was the flagellum glycosylation cluster, the members of *A. mediterranea* CF2 and strain *A. macleodii* 673 share the main LPS O-chain cluster (fig. 1). In this case, the similarity between the common (putatively exchanged) fGI was relatively low (ca. 95%), but much higher than the average of the two genomes (ca. 81%) which is suggestive of an ancient transfer and significantly posterior to the split of the two species. *E. coli* and *S. enterica* have been shown to share O-antigens, but again the genes show low similarity consistent with ancient transfers (Reeves 1993). These examples illustrate the way in which cells can diverge in glycotype by replacing the gene clusters that code for the biosynthesis of these exposed structures. Milkman and co-workers suggested that the enormous diversity of O-antigens in *E. coli* derives from the capture of exotic clusters coming from distantly related microbes in “isolated and improbable exchanges” (Milkman 1997). This idea of “retransmission” of exotic gene clusters within a species was justified by the drive of escaping host immunity, i.e., they provide antigenic novelty that would be strongly selected for and will be preserved in the species. Along these lines, by searching with all the glycotype gene clusters of *Alteromonas*, we have found examples of syntenic clusters (albeit at very low similarity) in distantly related microbes. Specifically, a similar O-chain cluster (ca. 60%) to the one of *A. macleodii* EZ55 was found in *Catenovulum agarivorans* YM01, an Alteromonaceae isolate from the Yellow Sea (Shi et al. 2012) (supplementary fig. S6, Supplementary Material online). We have found also that the flagellum glycosylation cluster of *A. australica* H17 have a syntenic cluster in *Glaciecola polaris* LMG 21857 (supplementary fig. S6, Supplementary Material online). These examples illustrate that glycotype clusters are recruited from a pool of microbes of taxonomic diversity at least at the rank of family. However, it is important to emphasize that *Alteromonas* is a free-living microbe never exposed to any immune system.
Additive fGIs

Most synteny breaks correspond to additive fGIs that contain different gene cassette combinations, although their functions are sometimes related (fig. 5). The simplest fGI (and also the most common) are integration hotspots, most (6 out of 8) have tRNA and tmRNA genes at the insertion site. The additions of some cassettes were still hall marked by the tell-tale repeat of part of the tRNA gene downstream of the inserted cassette (direct repeat). Nearly all the strains have different versions of these fGIs except for those belonging to the same CF, and even those sometimes differed by several cassettes (López-Pérez et al. 2013). The presence of different cassettes at the same location within the same CF attests to their change in short evolutionary times. The number of additive fGIs was significantly higher in the right replichore (fig. 5) that is also enriched in tRNA genes.

Additive cassettes were obviously acquired by HGT as shown by their disconnection to the core phylogeny, i.e., the number and order of the shared cassettes is not dependent on the phylogenetic position of the strain, even when this phylogeny indicates very close last common ancestor (see below). Actually, the most remarkable finding when comparing additive fGIs of different species was the frequent detection of nearly identical cassettes. Figure 6 shows an example of cassette conservation in the heavy metal resistance additive fGI of the genus Glaciecola.

![Figure 6](https://example.com/f6.png)

**Fig. 6.**—Schematic representation of the tRNA-Phe fGI gene cassettes found in different strains representatives of different species of *Alteromonas*. The conservation of some cassettes present in widely divergent genomic backgrounds is apparent. Some cassettes were found still syntenic but at lower similarity in members of the genus *Glaciecola*.
Intraspecies Diversity is Key to Bacterial Evolution

The main change in pace in genomic variation appeared at the level of different clones or strains within the same species, i.e., over the 95% ANI threshold. These strains show strong evidence of recombination along the core and present multiple variations in their flexible genomes. They have different glyco-type determinants, which provide them with different surface properties and different targets for phage recognition. In Alteromonas, a few examples of shared glycoype islands reveal how these critical gene clusters are exchanged with other clonal lineages sometimes outside of the species boundaries. However, as shown by their stability within CFs, the exchange of these islands is rare enough to keep their linkage with the main determinants of the lineage ecological properties, at least within ecological time frames (López-Pérez et al. 2013; Kashtan et al. 2014), preventing their sweep across populations as some authors have claimed (Cordero and Polz 2014; Takeuchi et al. 2015). The exchange between species appeared to be older (containing more SNPs) than those within species indicating that these exchanges are probably even much less frequent, but they might be essential to renew the stock of glycosydic receptors available within the species. This phenomenon has been known for long in E. coli and Salmonella (Reeves 1993) and might be relevant for antigenic variation in pathogens with large environmental reservoirs such as Vibrio cholerae (Mutreja et al. 2011). Additive fGIs were the most obvious drivers of intraspecies diversity and represent the largest part of the flexible genome. Actually, their gene cassettes seem to flow rapidly through taxonomic entities of at least the rank of families and provide adaptive traits that make each CF different at multiple levels. Those include, metabolism (transporters, degradation pathways, efflux pumps, porins), environmental sensing (two component systems) protection from abiotic (metal resistance) and biotic factors (antibiotic production and resistance, CRISPR, toxin-antitoxin, restriction modification systems) (Gonzaga et al. 2012; López-Pérez et al. 2012, 2013; López-Pérez, Gonzaga, et al. 2014).

By contrast, the comparison of genomes belonging to different species did not bring much more diversity into the picture. Actually, the numbers of new genes found when comparing two strains of A. mediterranea and A. macleodii did not increase significantly over the figures found among CFs of each species. Even dN/dS values increased little, indicating that positive (Darwinian) selection did not increase when considering different species, i.e., there seemed to be little increase in the adaptive evolution of the core. The main difference found in the comparison between strains of different species was that recombination along the core seemed to be much less prevalent. This supports a decline in gene flow correlated with core genome divergence that, as has been suggested before, might lead to prokaryotic speciation (Shapiro et al. 2012). However, it is important to emphasize that the exchange of additive cassettes must happen frequently among different species, as shown by the detection of identical cassettes in distantly related ones. This proves that there is no genetic isolation of the different genospecies.

A decrease in recombination was observed when analyzing the genomes of Vibrio cypriolyticus associated to different particle sizes (Shapiro et al. 2012). Something similar could be taking place in Alteromonas and actually, the description of the genomes of A. mediterranea DE and A. macleodii ATCC 27126 revealed differences consistent with adaptation to particles of different size. Alteromonas is a typical bloomer that relies on fast growth rates to compete in marine oligotrophic waters, which probably restricts its habitat to the relatively nutrient rich and extremely diverse realm of particulate organic matter. In fact, typically pelagic free-living microbes such as Prochlorococcus, Ca. Pelagibacter or the ammonia oxidizing archaea seem to have less species diversity (Penno et al. 2006; Gilbert et al. 2008; Hatzenpichler 2012).
Niche separation would lead to different phenotypic traits that would fit the classic polyphasic definition of bacterial species (taking into account all available phenotypic and genotypic parameters) (Ramasamy et al. 2014). Another factor to consider would be the barriers to recombination and exchange derived from the decrease in co-infecting phages, i.e., much fewer phage lineages can infect members of two separate species (Dixit et al. 2015) as a consequence of the divergence and lower exchange rates of replacement fGIs coding for cell surface phage targets.

A Model for Pangeneome Expansion, from Clones to Species

The ecotype speciation theory sustains that each ecotype exploits a specific niche. Genetic heterogeneity should be purged from the population by periodic selective sweeps (Cohan 2002; Lassalle et al. 2015). However, different lines of evidence from metagenomics, single cell genomics and culture of concurrent clones indicate a large diversity in different populations of aquatic microbes (Gonzaga et al. 2012; Kashan et al. 2014; Bendall et al. 2016). We would like to posit that the survival of a single clone in nature is unlikely because the restricted genetic information carried by one prokaryotic cell might not be enough to allow the population to adjust efficiently to environmental changes, even in one single niche (e.g., particulate matter in the ocean water column). In addition, a monoclonal population could be preyed to extinction by specific phages (Leggett et al. 2013). On the other hand, the coexistence of several clones with different ecological attributes not only dilutes the selective pressure of phages, it can be good for the population as a whole that can exploit resources more efficiently and share the load of interaction with stressful environmental factors as suggested by the Black Queen hypothesis (Morris et al. 2012). This equilibrium can be maintained with a complement of phages that prevent any single clone from replacing the others (a clonal sweep or periodic selection event). This model was proposed as “constant-diversity” dynamics (Rodriguez-Valera et al. 2009) and the metaclonal population including the phages proposed to be a single selection unit (Rodriguez-Valera et al. 2009; Rodriguez-Valera and Ussery 2012).

With these hypotheses in mind and with our set of strain genomes, it is possible to visualize how this high clonal diversity could be established (fig. 7). As a clone multiplies asexually, it diverges at the level of SNPs as both mutation and recombination take place along the core. At the same time, genomes incorporate different additive cassettes received from other microbes by HGT. This way clonal lineages start to diverge and acquire different traits. We can see how this is starting to happen in the populations represented by the five isolates in CF1. However, as long as they have the same replacement genomic islands (belonging to a single glycotype), phages prey similarly on all the cells, and there is the risk that a clonal sweep or a massive lytic event will make the incipient clone extinct. On the other hand, if one clone manages to persist for long enough, it will eventually acquire a different set of replacement fGIs (phage receptors) becoming a different glycotype, as seems to be starting to happen in strain U4 (fig. 1). The new glycotype with the cognate set of phages (that could be transplanted from the same origin as the glycotype determinant), will then become integrated in the metaclonal population. Incidentally, phage receptor swapping could increase recombination rates by originating phages with broad infection range that can transfer genes between more different genomic backgrounds. Understanding the nuts and bolts of prokaryotic evolution will require profound changes in our way to think about population genetics and evolution at large. Comparative genomics provides a wonderful tool to start unraveling its intricacy.

![Fig. 7](https://example.com/fig7.png)
**Supplementary Material**

Supplementary tables S1 and S2 and figs. S1–S6 are available at Genome Biology and Evolution online (http://www.gbe.oxfordjournals.org/).

**Acknowledgments**

We are grateful for the suggestions to improve the manuscript of some of the referees during the review process. This work was supported by projects MEDIMAX BFPU2013-48007-P from the Spanish Ministerio de Economı´a y Competitividad, MaCuMBA Project 311975 of the European Commission FP7 and PROMETEO II/2014/012 project AQUAMET from the Generalitat Valenciana. Strain D7 was kindly provided by Professor Åke Hagstro¨m (Linnaeus University, Sweden).

**Literature Cited**


Associate editor: Eric Bapteste