OPINION

Explaining microbial population genomics through phage predation

Francisco Rodriguez-Valera[1], Ana-Belen Martin-Cuadrado[1], Beltran Rodriguez-Brito[2], Lejla Pasic[3], T. Frede Thingstad[4], Forest Rohwer[2], Alex Mira[5]

[1]División de Microbiología, Universidad Miguel Hernández, Campus de San Juan, Alicante, Spain

[2]Department of Biology, San Diego State University, California, USA

[3]University of Ljubljana, Department of Biology, Ljubljana, Slovenia

[4]Department of Biology, University of Bergen, Bergen, Norway

[5] Área Genómica y Salud, Centro Superior de Investigación en Salud Pública, Valencia, Spain

Abstract

The remarkable diversity of genes within the pool of prokaryotic genomes belonging to the same species or pan-genome is difficult to reconcile with the widely accepted paradigm which asserts that periodic selection within bacterial populations would regularly purge genomic diversity by clonal replacement. Recent evidence from metagenomics indicates that even within a single sample a large diversity of genomes can be present for a single species. We have found that much of the differential gene content affects regions that are potential phage recognition targets. We therefore propose the operation of Constant-Diversity dynamics in which the diversity of prokaryotic populations is preserved by phage predation. We provide supporting evidence for this model from metagenomics, mathematical analysis and computer simulations. Periodic selection and phage predation dynamics are not mutually exclusive; we compare their predictions to indicate under which ecological circumstances each dynamics could predominate.

Explaining patterns of diversity in microbial populations has been one of the great conundrums of Microbiology. Historically, the approaches used for studying bacterial population genetics have been derived from eukaryotic models. They were based on distinguishing alleles and quantifying their presence in populations, as well as their degree of linkage, to infer mutation rates and recombination (see for example [1]). However, with the advent of genomics, it has become apparent that the genomic diversity in prokaryotes derives much more from having different sets of genes than by allelic differences at the same loci [2-4]. This is in contrast to eukaryotic organisms that preserve remarkably well the gene content within the same species and even across large phylogenetic distances [5]. The concept of the pan-genome has been coined to describe the increasing diversity of the gene pool that can be ascribed to one bacterial species as the number of sequenced strains increases [6].Typically, the comparison of any couple of strains might reveal about 10-35% of the genome content (typically in the range of 500-1000 genes) that is present in only one of the strains but not in the other [7] even though the comparison of the core-genome regions may indicate that they are highly related lineages belonging to a single species. The implications of these findings are still permeating into the scientific community but their importance for our way of thinking about prokaryotic microdiversity and evolution is paramount. Is it relevant important for the ecology and environmental adaptation of the different lineages or just the result of junk DNA accumulation that has yet to be pruned by regular sweeps of natural selection? What are the evolutionary forces that preserve this degree of diversity within highly related populations?

The accepted models of bacterial population genetics sustain that in asexual microbial populations a low phenotypic diversity is expected because of purges involving fitter mutants, called periodic selection events [8]. This idea originated from classical laboratory experiments of mutational equilibrium in which populations are periodically replaced by new types [9, 10]; it was later reinforced by epidemiological studies of pathogenic isolates in which a rise and fall of clonal lineages do occur [11, 12].This process would purge diversity from the population, at least among the cells competing for the same resources. The same kind of dynamics has also been claimed to be behind the genetic coherence of natural prokaryotic taxonomic units or “ecotypes” [13, 14]. Ecotypes are defined as “populations that are genetically cohesive and ecologically distinct”. Cohesion is mostly ascribed to “periodic selection events that recurrently purge each ecotype of its genetic diversity” [15, 16]. Divergence can become permanent when a mutation (or recombination) event places an organism into a new ecological niche and founds a new ecotype [13]. Periodic selection will therefore keep the

populations within ecotypes relatively homogeneous and divergent from other ecotypes. In the Stable Ecotype Model an ecotype is recurrently purged of its diversity by these selection events, whereas divergence among ecotypes is not [16] becoming thus “the fundamental units of ecology and evolution”.

Presently, the existence of natural diversity units of bacteria or ecotypes is widely accepted. However the origin of such units by regular periodic selection events seems difficult to reconcile with the wide gene content variability found among different strains of bacteria with otherwise high sequence conservation among shared genes (i.e. the pan-genome). We know that prokaryotes can easily acquire foreign DNA, and phages and plasmids can be easily transferred and even inserted into the main chromosome. In fact, some (or most) periodic selection events could be due to the acquisition of novel genes rather than mutation. However, the neutral accumulation of a large number of genes within a clone before a clonal sweep purges it seems unlikely, and that phenomenon might have alternative explanations.

We propose that the main factor, which has been largely overlooked in many previous models, is the role played by a crucial ecological factor: the presence of bacteriophages. In nature, prokaryotic cells have to deal with a strong predation pressure mainly viral in origin and therefore their fitness is measured not only by adaptation to the physical niche but also by adaptation to the biotic environment [17]. Protozoan grazing might also contribute [18] but it is less pervasive than viral attack [19].

In almost all ecosystems that have been investigated, there are ~10 phages for every microbial cell, making them the most abundant biological entities on the planet [20]. By killing microbes, phage greatly influence global biogeochemical cycles and because phage are species-specific they have been predicted to help maintain microbial species diversity [21, 22]. In fact, phages could play a fundamental role as guarantors of the microdiversity required to exploit ecological resources efficiently. We propose a new type of dynamics to explain bacterial microdiversity which incorporates phage predation. We begin by systematically describing the content of variable genomic regions among closely related bacterial strains with the aim of identifying the force that drives microdiversity. We then introduce a Constant-Diversity (C-D) dynamics to explain the generation and maintenance of microbial diversity in natural ecosystems. We compare its predictions from those of the Periodic-Selection (P-S) and evaluate the evidence based on mathematical modeling, metagenomic data and microbial ecology studies. Finally, we reconcile the C-D and P-S dynamics on the basis of the ecological features of different environments that would favor either one.

Single species metagenomics

Metagenomic approaches have demonstrated the value of comparing genomes of individual strains with the metagenomes from environments in which these species or ecotypes are present. Such comparisons have shown that certain genomic regions are underrepresented within any given metagenome and are therefore predicted to be unique to individual isolates [23-29]. The metagenome represents all lineages within one sample and, if it has enough coverage, it should contain representatives from the predominant ecotypes. This virtual experiment has actually been carried out a number of times [24-30], showing regions that recruit poorly in the metagenome. These underecruiting genomic regions or Metagenomic Islands (MGIs) are therefore predicted to be unique to individual cell types. It is important to note that these MGIs must not be confounded with classical genomic islands, i.e. regions of unusual DNA composition that are generally indentified with lateral gene transfer, and have typically the hallmarks of mobile DNA elements or recombinatorial hot-spots. Nevertheless, some MGIs are obviously of foreign nature and show extraneous compositional features. One ideal system for this kind of comparison is provided by extreme environments that are heavily dominated by few microbes (very low diversity). For example, a saturated brine that is largely dominated by the halophilic archaeon *Haloquadratum walsbyi* [24]. When the brine metagenome was compared to an already available genome of the strain isolated from the same pond many genes were either extremely divergent, highly rearranged or not found in the metagenome [27]. They included large numbers of genes encoding the cell surface components, which in the case of *Haloquadratum* are glycoproteins, as well as glycosilation of surface components (Figure 1) [27]. Many genes in the variable pool were the sensors of two component regulators and substrate transporters. These two features were interpreted to indicate differential adaptation to phage sensitivity and organic carbon degradation. A very similar pattern was found when again a nearly monospecific metagenome dominated by *Candidatus* Accumulibacter phosphatis was sequenced in sludges from enhanced biological phosphorus removal reactors [30]. One could argue that these are special cases, since a single species has to perform many metabolic roles that would be shared among many in high diversity environments. However, a strikingly similar pattern of inferred functions is found even within metagenomic islands from high diversity environments like the open ocean (Figure 1).

We have systematically studied MGIs larger than 10 kb in all sequenced marine bacterial species compared to a marine metagenome. The content of these MGIs was

strongly biased with respect to specific functional categories (Supplementary Figure 2 and Supplementary Table 1). It is remarkable that all studied species contain within these MGIs genes that encode products that are extracellularly exposed (Figure 1). Paramount among them we find the variable O-chain of the LPS, which has long been known to be highly variable and a choice target for phage receptors (see, for example, [31]). Serotyping and phage typing are usually linked to O antigen changes determined by this gene cluster [32]. In pathogens this variability has been often explained as host-immunity evasion strategies, but in free-living microbes other explanations are required. Next by frequency are exo-polysaccharide biosynthetic clusters and/or sugar decoration of extracellular structures. Pilli and flagellar components (particularly their extracellular components) are also found. Recently, [33] an island was found containing an alternative set of external flagellar proteins. MGIs also frequently harbour giant proteins, whose function is still unclear, but that are probably extracellular [34] and porins. All those genes are potential phage recognition sites suggesting a potential role in phage avoidance. When the functional classification of genes encoded within MGIs is compared to the proportions found in the genome, the genes coding for potential phage recognition sites are heavily over-represented, followed by genes involved in nutrient transport and environmental sensing (Supplementary Figure 1). Thus it can be concluded that the overall phage-interacting genes tend to be non-shared and when they are shared, they tend to be more highly divergent. Phages depend heavily on the proper selection of the target cell and for that they rely on a prominent structure for their target recognition [35, 36]. It is important to note that the high divergence in potential phage-recognition sites is found even in extremely compact genomes like *Candidatus* Pelagibacter ubique (Figure 1), which has the smallest sequenced genome among marine prokaryotes. Despite its compact genome, the islands in the three available strains of Pelagibacter contain surface features (Figure 1a) and transport and sensing genes. Thus, the presence of potential phage targets in the strain-specific areas of the genome is a pervasive phenomenon in the open ocean, suggesting that this feature may apply to other free-living prokaryotic microbes subject to phage predation pressure. The second line of phage defense is intracellular. Our MGI data has numerous examples of these types of signatures such as restriction-modification systems and the clusters of regularly interspaced palindromic repeats (CRISPR) that are involved in phage interference [37-39]. However, they are probably involved in infection efficiency providing a fine-tuning that would reinforce the primary control exerted by the receptor diversity.

Constant-Diversity dynamics

The C-D dynamics is intended to explain microbial diversity in ecosystems where bacterial populations can interact with each other, i.e. the populations must compete with each other and phage particles have similar chances to infect any cell within the community. Let's imagine an idealized aquatic habitat where organic nutrients are dissolved and in which a single prokaryotic species or ecotype is present. A large diversity of phage-sensitivity-types is required to avoid catastrophic lysis of the population. We will call these phage-receptor-types R1, R2…Rn (R stands for phage receptors), each recognized by a phage lineage, denoted as P1, P2…Pn. The phage-types differ at the level of gene clusters that control the synthesis of complex surface components such as the O-chain or the pilli, which are extracellularly exposed and act as targets for phage recognition, analogous to a lock and key system. P-S occurs when a new adaptive mutant (or recombinant) arises within the ecotype and natural selection causes the mutant and its nearly clonal descendants to replace all competing variants within the ecotype [15]. However, we predict that the increase in number of that fitter lineage would alter the predator-pray equilibrium and phages targeting the receptor coded within this lineage would also increase (Figure 2a). This would select against the invasive clone that would eventually be replaced by the original "normal fitness" lineages. This way a constant high-diversity of lineages would be maintained steadily. Interestingly, the role played by phage predation modifies the classical 'survival of the fittest' axiom such that metabolically superior microorganisms which are better adapted to a physical environment are selected against by the biological pressure imposed by density-dependent phage predation. This "kill-the-winner" dynamics have already been proposed for different species of marine bacteria [21, 22] but here we propose that this process is responsible for maintaining closely-related lineages diversity. In a C-D situation, no dominant lineage is found within a population, and phage dynamics maintain many concurrent cell types that are selected against ecological success. This success will be influenced by factors such as growth rate or the efficient use of different substrates and therefore a corollary is that throughout the history of these clonal lineages each will acquire different, complementary capabilities for niche exploitation. As a consequence, a more efficient exploitation of the resources by the community is expected and a better ecosystem functioning would be achieved. For example, one prokaryotic clone cannot contain even a fraction of the transporters required to internalize the chemical diversity of organic compounds contained in a single eukaryotic cell. However, an ensemble of lineages carrying different sets of transporters could exploit every single one of them**.** A relationship between biodiversity and ecosystem efficiency has been demonstrated in plant ecosystems [40] and we predict that similar principles would apply for biogeochemical cycles controlled by

microbial communities, in such a way that ecosystem functioning would be more efficient and stable under high-diversity situations than under periodic clonal sweeps, as the latter would make nutrient and mineral recycling fluctuate (Figure 2b).

Computer simulations of bacterial strain replacement in the presence and absence of phages show that viral predation affects biodiversity (Figure 3). In the absence of phages, microbial density for each lineage is dependent on the availability of the substrates and the efficiency of nutrient utilization. Thus, the lineage that uses the most abundant substrate at a given time is the most abundant (Figure 3c). If an environmental change (e.g. a variation on the concentration of nutrients) or a mutation (e.g. a change in nutrient utilization efficiency) occurs, the fitter lineage will be favoured, generating a clonal sweep. As a consequence, the diversity of the ecotype will be low (Figure 3b), as the better adapted strain dominates over other strains with smaller representation in the consortium. In the presence of phages, however, a richer lineage that uses the most abundant nutrient and/or that is able to metabolize more than one substrate will be preferentially attacked by phages, because phage-host interactions are density-dependent [21]. This way the fitter bacterial strains are selected against, thus the density of the different bacterial lineages fluctuates around stable levels (Figure 3d). Therefore, bacterial diversity remains high, with all lineages present at similar values regardless of the availability of the substrates they utilize (Figure 3c). The immediate consequence of this fine-tuning to different substrates would be the expansion of the gene pool within the ecotype. Thus, contrary to what would be expected by clonal replacements, C-D dynamics would predict a large pan-genome within ecotypes (Box 1).

Another interesting feature shown by these simplified simulations (Figure 3d) relates to ecosystem functioning. In the presence of phage predation pressure and environmental changes, the lineage feeding on the most abundant substrate (or utilizing efficiently more than one substrate) undergoes substantial fluctuations in density. Lineages utilizing the least abundant substrates, on the other hand, show constant densities through time, undergoing only slight variations. Because the amount of density fluctuations is directly related to the probability of extinction, selection will favour less fit cells utilizing single, low-concentration substrates. This paradoxical selection against the fitter cells will give rise to lineages feeding on all accessible substrates, regardless of their availability. As a consequence, the ecosystem is expected to be more efficient. This high efficiency in substrate utilization at the ecosystem level will be sustained by the use of all available metabolites and minerals and by the presence of a high diversity, which is known to promote efficient ecosystem functioning [40, 41]. Note that a

high bacterial diversity cannot be the consequence of the availability of different resources alone: it is only in the presence of phages when all cell types reach a similar average density regardless of the substrate they utilize; in the absence of phage predation, the exploitation of less favoured niches is selected against.

Predictions from Constant-Diversity dynamics

Despite several refinements all the Ecotype based models rely strongly on the relevance of P-S based on competition for the resources among individual lineages (see for example [15]). C-D on the other hand predicts that P-S of an individual lineage with larger capabilities to exploit resources would be prevented by phages, and a better exploitation of resources is achieved by many different lineages sharing the environment. Both views make quite different predictions about microbial biodiversity and population dynamics (Table 1). Under P-S, a dominant cell type would be generally found, and this dominant lineage would change periodically. Under C-D dynamics, on the other hand, phages would keep in check dominant lineages and therefore many coexisting cell types would be found at any time. In addition, the former predicts that genomic variability among ecotypes is driven by niche exploitation and the latter that it is driven by phage avoidance (Table 1).

What is the evidence for each of these models? The coexistence of multiple closely related strains is a commonly observed phenomenon in microbial communities, and is apparent from studies of cultured isolates [42], marker gene surveys [43], and metagenomic data [27, 44] (the ecological diversity of close relatives has been discussed in [15, 45]). In addition, mathematical modelling of bacterial dynamics in the presence of phage indicates that a dramatically high number of strains is expected, as observed in metagenomic data (Box 2). Although a high number of strains has indeed been observed in free-living ecosystems, the maintenance of this high diversity through time remains to be explained. The evidence for phage-bacterial antagonistic dynamics first came from mathematical modelling as well as experimental studies [21, 22, 46, 47]. Recently, excellent work on marine samples has shown that phages do influence their bacterial hosts in a density-dependent manner, mainly infecting a reduced number of phylotypes at any one time [48], as predicted by kill-the-winner dynamics. Similar results have been found in the horse gut when studying the relationship between coliphages and *E. coli* strains [49]. In addition, detailed studies on marine flavobacteria have shown the degree of phage susceptibility in different co-inhabiting cell types, drawing a complex picture of phage-cell interactions [50]. Recent data on the temporal variation in phage and bacterial diversity also show oscillations that are consistent with a constant control of abundant genotypes by their infecting phages (Rodríguez-Brito *et al*., under review): the data

show that dominant genotypes are not found through time and that a dynamic equilibrium of functionally-redundant microbial and viral strains continuously replaces each other in a Kill-the-Winner fashion, thus maintaining a stable metabolic potential and taxonomical signal. In terms of resource use, there is no experimental evidence showing that some bacterial cells within an ecotype are more generalist or have superior fitness. Furthermore, mathematical modelling indicates that evolution of microbial systems does not lead to decreased biodiversity by the expansion of a variant that is ecologically superior [51]. On the contrary, it has been shown that some co-existing strains differentially specialize in micro-niches thanks to their different gene content, for example by utilizing different sets of organic compounds heterotrophically [27] or by exploiting different types of particulate matter [33]. This supports the notion that selection favours resource diversification rather than a dominant genotype capable of using multiple substrates. It would nevertheless be interesting to experimentally measure the individual fitness of different bacterial strains isolated from nature in order to determine whether fitter variants are selected against in environments under phage pressure.

Other predictions are more difficult to evaluate with the available data. Ecosystem functioning, for example, has been thoroughly studied in terrestrial ecosystems, where a high biodiversity has been related to resource utilization efficiency [40, 41]. This information is still limited in microbial ecosystems but they also relate microbial diversity to ecosystem efficiency and stability [52]. It would be interesting to quantify the efficiency and stability of ecosystem functioning in natural conditions and in the laboratory where both bacterial diversity and the presence of phages can be manipulated in microcosm experiments [53, 54]. We would predict that the presence of bacteriophages would promote resource specialization in their bacterial hosts, regardless of the respective availability of substrates, facilitating efficient nutrient recycling and system stability.

The C-D dynamics counter-intuitively predicts that the strain of higher fitness would be selected against and that phage predation would select for cell types of lower fitness that utilize any substrate regardless of its availability (Figure 3). By keeping a low profile in the population, the cells would be part of a selection unit formed by bacterial strains of sub-optimal but nearly-equal fitness where predation risk is equally shared among the strains. Thus, the fitness concept under P-S dynamics is dependent on the efficiency with which a resource is used but the evolutionary success of a strain under C-D dynamics is dependent on the fitness of the other strains co-inhabiting the

environment and is therefore a relative concept. This is analogous to the outcome of game theory models for optimal market strategies: there is no optimal economic strategy *per se*; instead, the best market tactic depends on what the competitors are doing. This thinking has been successfully applied to evolutionary theory, giving rise to models that conclude that once an optimal proportion of strategies is achieved in a population, those proportions are in equilibrium over time [55]. We believe these Evolutionary Stable Strategies (ESS) best define the situation in bacterial populations in the presence of phages, because under those conditions the fitness of bacteria would be lowered to an equilibrium point in which demographic success would be approximately equal among cell types. Any strain increasing its resource utilization would be quickly eliminated by phage predation and an ESS would necessarily imply adjusting growth rates to those of the rest of the population, unless the cell is immune to phage attack for some reason. The idea does not go against natural selection but it does require that bacterial "fitness" includes ecological aspects such as predation avoidance [17].

Microbial Evolution under phage pressure

Under which circumstances should we thus expect P-S or C-D to predominate? The influence of phage predation on bacterial diversity requires that bacterial populations interact with each other; therefore host-associated niches can act as physical barriers preventing direct cell competition and phage dispersal. Thus, pathogens are not expected to follow C-D dynamics unless they become very numerous and persist within a host for a long time. C-D dynamics are also not expected in physically-constrained microbial communities such as biofilms, where populations cannot interact with each other nor invade other niches, apart from constraining phage attack and dispersal. By definition, C-D dynamics require the presence of phages and therefore are expected to be less common in specific environments with limited viral presence or efficiency. Accordingly, large MGIs are not observed in biofilms as the work carried out in the acid mine drainage system shows [56]. Thus, although the presence of C-D patterns mediated by phage predation is pervasive in natural environments inhabited by competing, free-living cells, the two population dynamics are not mutually exclusive and P-S selection is still expected to occur in circumstances of low predation pressure (e.g. intracellular environments), physically-constrained cells (e.g. biofilms) or physically isolated environments subject to founder effects (e.g. animal hosts).

The C-D dynamics presented here may shed some light on different aspects of microbial evolution. One of them is the function of giant proteins, which are over-

represented in metagenomic islands and that could therefore be involved in host-phage interactions. In addition, if predation pressure is so intense, we would predict that genes providing phage resistance would be among the fastest-evolving in free-living species in order to counteract the fast phage adaptability [57], giving rise to an evolutionary arms-race [35] analogous to the one between the immune system and surface antigens in bacterial pathogens. Another aspect that should be explored is the potential role of phages in defining the limits of bacterial populations or ecotypes. In that sense, phage infection may provide a link between the evolutionary process of generating diversity and the mechanistic explanation for it, which indicates that lack of recombination is responsible for population genetic divergence [58]. The phages themselves could help homogenization of sequences by recombination, by providing a constant flux of homologous sequences among the population that is co-infected by the same phage type. With the sequencing of viral metagenomes [59], our understanding of the diversity and specificity of phages will be improved, and mathematical modelling of the viral-bacterial interactions should become possible under realistic assumptions. In addition, experimental evolution studies in presence and absence of phages [17] and further genomic characterization of natural environments with different phage pressure should confirm whether the P-S and C-D dynamics complement each other and may explain evolution of intraspecific microbial diversity under different ecological circumstances.

Acknowledgements

Work in the laboratory of F. R-V is supported by Grant BIO2008-02444 of the Spanish Ministry of Science and Innovation. A.B. M-C is supported by a Juan de la Cierva postdoctoral fellowship of the Spanish Ministry of Science and Innovation. A.M is funded by the CSISP from Generalitat Valenciana. Funding at F.R. lab was received from the US National Science Foundation (DEB-BE04-21955), Gordon and Betty Moore Foundation, ARCS Fellowship, NSF Postdoctoral Fellowship #DBI-0511948, ATP/Kent Sea Tech. We thank Dr. L. Pasic for help in the analysis of *S. ruber* genomic islands.

FIGURE LEGENDS

Figure 1. Metagenomic Islands (MGIs) identified by comparison with available metagenomes. Representative species were selected that recruited at >90% DNA sequence identity over >80% of the genome. (a) Genome recruitment of the marine alpha-proteobacterium *Candidatus* Pelagibacter ubique HTCC1062, HTCC1002 and HTCC7211 when compared with the Global Ocean Survey database (GOS, Phase I [60] and Phase II). Individual sequences were aligned to the sequenced strain genome and the alignment-sequence conservation visualized in the form of percent identity plot. Regions larger than 10 Kbs with unusually low representation in the metagenome (MGIs) are marked with a brief description of their main features. (b) Genome recruitment of *Prochlorococcus marinus* MED4, *P. marinus* MIT9301, *Burkholderia* sp. 383 chr.1, *Shewanella* sp. MR-4, *Aeromonas hydrophila subs. hydrophila* ATCC 7966, and *Synechococcus* sp. WH812 when compared with the GOS dabatase. The recruitment of *Salinibacter ruber* DSM 13855 and *Haloquadratum walsbyi* DSM 16790 was performed with solar saltern metagenomes [24, 61]. MGIs related to extracellular components are marked following the colour pattern indicated in panel (a). Recruitment plots were collected from http://gos.jcvi.org/openAccess/genomes.html with the exception of *S. ruber* DSM 13855 and *H. walsbyi* DSM 16790. A full description of the MGIs contents from all species that recruited over 80% can be found at Supplementary Table 1.

Figure 2. Population dynamics under (a) Constant-Diversity and (b) Periodic-Selection. Several sympatric clones with diverse capabilities to exploit environmental resources coexist within an ecotype in an aquatic habitat simultaneously. The different resources are represented by geometric symbols, indicating a different niche. These lineages differ in the set of receptors to phages, indicated by R#. Any strain from one specific lineage will recruit very poorly for the genes responsible for the R#, which would be clustered in metagenomic islands. Many genes throughout the genome and required for the exploitation of the niche are linked to the corresponding R# and hitchhike with it. At some point in time one lineage will gain in fitness and replace other lineages. This could be achieved by a mutation (including a Horizontal Gene Transfer event, indicated as HGT) or by an environmental change favouring one particular genotype. Under Constant-Diversity (a) this situation is only transient because an increase in the frequency of cells carrying the specific phage receptor would unsettle the predator/pray equilibrium and select against the invasive clone that would

eventually be replaced by the original “normal fitness” lineages. Kill-the-winner dynamics would therefore give rise to a constant, high diversity of lineages that would be maintained steadily. By these equilibrium a more efficient exploitation of the resources by the community is predicted and a better ecosystem functioning achieved. Under Periodic-Selection dynamics (b), lineages with higher fitness would expand and replace other cell types, giving rise to a clonal sweep. The process would be repeated after advantagous mutations or when environmental changes increase the fitness of a given strain, decreasing diversity. Ecosystem efficiency is also expected to be reduced, as the lineages exploiting substrates of low availability would severely decline.

Figure 3. Cell density and microbial diversity under Constant-Diversity and Ecotype dynamics. The graphs show the outcome of computer simulations from a simplified model of an ecosystem inhabited by six bacterial lineages, each of them utilizing optimally one specific substrate, whose availability varies through time (a). In the absence of phages (c), microbial density is directly related to the availability of the substrates and the efficiency of nutrient utilization. After an environmental change (e.g. a variation in nutrient concentration) or a mutation (e.g. a gene transfer improving nutrient utilization efficiency), the fitter lineage is favoured, generating a clonal sweep. As a consequence, the diversity of the system (b) measured by the Shannon Index in the figure, drops. In the presence of bacteriophages, the lineage of higher metabolic fitness is preferentially attacked by viruses, assuming density-dependent phage-host interactions. This kill-the-winner dynamics selects against the fitter bacterial strains, so the density of the different bacterial lineages fluctuates slightly (note the differences in scale) around stable levels (d) and bacterial diversity remains high (d), with all lineages present at similar values regardless of substrate availability. The model is a coupled set of differential equations and is composed by six pairs of equations which are repeated for each of the six substrates. There is no direct interaction between different strains of viruses, and the interaction between strains of microbes comes in the form of a common carrying capacity for the whole system. We have considered a simplified case in which one phage infects one cell type but it must be born in mind that several experiments indicate that host-virus interactions in natural systems are more complex, with each phage infecting different bacterial strains with different efficiencies [50, 62] (in other words, the binding probability between bacteriophages and cells is graded [63]). For the sake of simplicity, the decay rate was assumed to be the same for all the viruses. The same was true for the mass-action constant as well as the burst size. This

model was programmed in matlab (The MathWorks, Natick, MA, USA) and the code is available upon request.

Supplementary Figures

Figure S1. Relative frequencies of different functional categories in Metagenomic Islands (MGIs) compared to their frequency in the genome. Graphs show the percentage distribution in each COG category for *Candidatus* Pelagibacter ubique (data from the genomes of strains HTCC1062, HTCC1002 and HTCC7211), *Prochlorococcus marinus* (genomes of strains *P. marinus* MED34, *P. marinus* MIT9301, *P. marinus* AS9601 and *P. marinus* MIT9312) and *Shewanella* (genomes of strains *Shewanella sp.* MR4 and *Shewanella sp.* ANA3). The genomic frequencies were calculated without taking into account the COGs found in MGIs.

Figure S2. Relative frequencies of different functional categories in Metagenomic Islands. Graphs show the percentage distribution in each COGs category for *Candidatus* Pelagibacter ubique (data from the genomes of strains HTCC1062, HTCC1002 and HTCC7211), *Prochlorococcus marinus* (genomes of strains *P. marinus* MED34, *P. marinus* MIT9301, *P. marinus* AS9601 and *P. marinus* MIT9312), *Shewanella* (genomes of strains *Shewanella sp.* MR4 and *Shewanella sp.* ANA3), *Synechococcus* sp. WH8102, *Burkholderia* sp.383, *Aeromonas hydrophila subs. hydrophila* ATCC7966, *Salinibacter ruber* DSM 13855 and *Haloquadratum walsbyi* DSM 16790. **A.** Detailed COG distribution of the categories: Carbohydrate transport and metabolism, Amino acid transport and metabolism and Inorganic ion transport and metabolism in *Candidatus* Pelabibacter ubique. **B.** Detailed COG distribution of categories: Carbohydrate transport and metabolism, Amino acid transport and metabolism and Inorganic ion transport and metabolism in *Prochlorococcus marinus*. **C.** Detailed COG distribution of the categories: Signal transduction mechanism, Carbohydrate transport and metabolism, Amino acid transport and metabolism and Inorganic ion transport and metabolism in *Shewanella*.

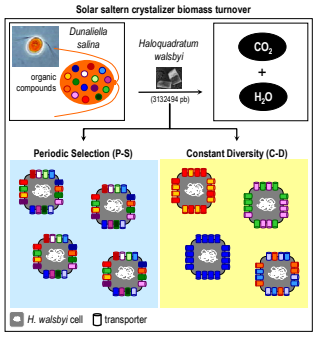
Box 1. Constant-Diversity and the Pan-genome

The nascent field of population genomics has shown that the gene pool of prokaryotic species can be extremely large. Different lineages of bacteria contain different genomes (similarly to the way the different tissues have different proteomes in a multicellular eukaryote) increasing enormously the metabolic and ecological capabilities of one bacterial species. As a consequence the concept of bacterial species has dramatically changed during the last years [7, 64] and bacterial species are more appropriately described nowadays by their "pan-genome", which includes a core-genome containing genes present in all strains and an accessory-genome consisting of partially shared and strain-specific genes. If a bacterial species is more than a semantic term and has in fact a biological meaning, the core (or "backbone") genome is the essence of this phylogenetic unit and is thought to be representative at various taxonomic levels [65]. The accessory (or adaptive) genome, on the other hand, includes key genes to survive in a specific environment, it is commonly linked to virulence, capsular serotype, adaptation and antibiotic resistance and might reflect the organism's predominant lifestyle [66]. Ecotypes have been claimed to have the quintessential properties of species i.e. they are ecologically distinct groups belonging to genetically cohesive and irreversible separate evolutionary lineages [15]. One key question is weather the pan-genome reflects the diversity of ecotypes within one species or, if actually ecotypes are the real bacterial natural species and a large pan-genome can be found again when sequencing different lineages within a single ecotype.

One of the arguments that is often used to explain ecotype genetic cohesiveness is the Periodic-Selection (P-S) dynamics, the genome-wide purging of diversity that occurs when a new adaptive mutant (or recombinant) arises within nearly asexual bacterial populations and causes the mutant and its nearly clonal descendants to replace all competing variants within the ecotype. If P-S occurs then the pan-genome of one ecotype would be very narrow. Constant-Diversity dynamics on the other hand would maintain the potential for a relatively large pan-genome within one ecotype by preventing any specific lineage from achieving a clonal diversity sweep. Although phage interacting genes are the part of the accessory-genome that is directly involved in C-D other adaptive genes hitch-hike with them allowing different lineages to have specialities within a single ecotype or genetically cohesive unit (the cohesiveness is clearly apparent in the core-genome). Although individual ecotypes can not be experimentally separated, simple environments in which ecological units are well defined can serve as a starting point to test whether a pan-genome structure exists within ecotypes. One such environment is the saturated brine dominated by a single prokaryotic species where most of the biomass providing nutrients comes from the degradation of the microalgae *Dunaliella* (see Box Figure). In this extremely halophilic, simple ecosystem, the bacterial cells that degrade *Dunaliella* form a coherent and distinct ecological unit that must approximate an ecotype. When the metagenome of 5 litres of water from this single-species ecosystem was sequenced, a surprisingly large

gene pool was found and the inferred pan-genome was twice the size of the sequenced isolate [24]. A large portion of this pan-genome structure was due to genes related to environmental sensing and nutrient transport [27] which is consistent with the idea that a genomically-diverse ecotype is needed to efficiently degrade all the organic matter, as the number of required metabolic and transport genes far exceeds those that can be squeezed within a single genome. Thus, single-species metagenomics support that a high diversity (or a large pan-genome) exists even within very limited spatial and temporal frames and are clearly in favour of C-D over P-S dynamics taking place, at least in homogeneous aquatic environments. Counter-intuitively C-D should enhance resource exploitation by maintaining a large genetic reservoir –the pan-genome- that can be used by the population. This has been recently shown in microbial microcosm experiments, as the evenness in community composition was shown to be a key factor in preserving the functional stability of an ecosystem [52].

Figure Box 1. Rodriguez-Valera *et al.*



BOX 2. How much richness can lytic viruses produce in a prokaryote community

The community of heterotrophic prokaryotes in the photic zone of pelagic environments is relatively stable, usually in the range 3 10^5 – 3 10^6 cells ml^{-1}. One model suggested for the mechanisms controlling richness within this community is that it is regulated by

lytic viruses [21]. Since this is a density-dependent loss mechanism, it will act selectively on host populations reaching a sufficient size. By preventing these from sequestering all the limiting resources and filling up the total community size of 3 10^5 – 3 10^6 cells ml^{-1}, viral lysis will leave room for other, potentially less competitive, host population to fill in the remaining part of the prokaryote community. The richness which such a mechanism can produce is the ratio between the size of the total community and the average size of the host populations. Assuming a steady state, the balance between production and loss of viruses of type *i* is given by the equation:

$(m_i - 1)\beta_i B_i V_i = \delta_i V_i$ Eqn. 1,

Where m_i, β_i, and δ_i are the burst size, the effective adsorption constant, and the specific decay rate, respectively, for the host-virus pair B_i ,V_i. The -1 comes from one virus being lost at infection. Eliminating V_i and solving for B_i gives the size of the steady state host population B_i:

$$B_i = \frac{\delta_i}{(m_i - 1)\beta_i} \qquad \text{Eqn. 2}$$

The highest richness allowable corresponds to the smallest possible values for all B_i, and thus to low δ_i, but high m_i and β_i – values.

A classical determination of the value for the adsorption constant β is given in [67] with a value of 0.24 10^{-8} ml min^{-1} $virus^{-1}$ obtained both experimentally and theoretically from diffusion estimates. This is expected to be a high value since e.g. defense mechanisms in the host could lower the effective value relative to the collision frequency based on diffusion theory.

Combining this with a, for natural systems, relatively high burst size of 100 viruses released per lysis and a relatively slow decay rate of 1 $week^{-1}$, Eqn. 2 gives ca 400 cells ml^{-1} as a low estimate for the host abundance that can sustain a population of free lytic viruses. For a community size of 10^6 ml^{-1}, viral lysis can therefore be argued to allow a richness of around 2500 different hosts.

Note that these “hosts” do not need to be different “species”. A host in one group able to construct a defense against one virus through e.g. incorporating a part of the viral genome in its CRISPR system [38] would potentially form a new host group without any other change in its genetic composition. Host groups in the sense used in this model could thus well be indistinguishable by methods such as e.g. 16S rRNA sequencing.

Table 1. Predictions from the Periodic-Selection and Constant-Diversity dynamics

Periodic Selection	**Constant-Diversity**
Low diversity of cell types, with a dominant ecotype	Many concurrent cell types
Abrupt changes in dominant cell type	Stable, high diversity of cell types
Dominant lineage is more fit in exploiting resources	Each different lineage is suboptimal in exploiting resources
Dominant lineage is more generalist	Resource diversification among strains
Ecosystem expected to be less efficient due to lack of resource specialization	Ecosystem expected to be more efficient due to resources use shared by many lineages
Dominant lineage changes through time	Different lineages are stable
Absent/limited phage pressure	High phage pressure
Variability among lineages is small and mainly neutral	Variability among lineages is large, adaptive and related to different sensitivity to phage predation and resource exploitation
The ecotype has a restricted pan-genome	The ecotype has a large pan-genome
Physically constrained populations (i.e. hosts, biofilms)	Interacting populations (i.e. free-living)
Species coherence given by regular clonal sweeps	Species coherence given by phage specificity
Competition driven by resource use efficiency	Competition driven by Evolutionary Stable Strategy

GLOSSARY

Metagenome: The total genetic repertoire of a given environment. It is formed by the pool of genes belonging to all the strains from each prokaryotic species inhabiting it.

CRISPR: Clustered, Regularly Interspaced Short Palindromic Repeats. Widespread genetic system in Bacteria and Archaea that consists of multiple copies of palindromic repeats. Flanked by the repeats there are short spacers of phage origin that provides acquired resistance against viral infection.

Kill-the-winner: Population dynamics of phage-bacteria interactions which postulates that an increase in a host population (winner) is followed by an increase in its corresponding phage predator that will increase its killing rate. It is analogous to classical Lotka-Volterra dynamics to explain predator-prey population dynamics.

MetaGenomic Island (MGI): Genomic regions found in many prokaryotic genomes that show absent or very limited representation in the DNA pool of the environment they inhabit. They can be identified because the genes within a MGI are under-represented in the metagenome. Metagenomically defined islands must not be confounded with islands of unusual DNA composition that are generally explained as lateral gene transfer events. Nevertheless, some metagenomic islands are obviously of foreign nature and show extraneous compositional features.

Pan-genome: The total gene pool of a prokaryotic taxon. It is formed by the addition of all genes found in the different strains from a given species (species pan-genome) or from an ecologically distinct population (ecotype pan-genome).

REFERENCES

1. Gupta, S. & Maiden, M.C. Exploring the evolution of diversity in pathogen populations. *Trends Microbiol* **9**, 181-5 (2001).
2. Whitaker, R.J. & Banfield, J.F. Population genomics in natural microbial communities. *Trends Ecol Evol* **21**, 508-16 (2006).
3. Medini, D. et al. Microbiology in the post-genomic era. *Nat Rev Microbiol* **6**, 419-30 (2008).
4. Medini, D., Donati, C., Tettelin, H., Masignani, V. & Rappuoli, R. The microbial pan-genome. *Curr Opin Genet Dev* **15**, 589-94 (2005).
5. Liti, G. et al. Population genomics of domestic and wild yeasts. *Nature* (2009).
6. Tettelin, H. et al. Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome". *Proc Natl Acad Sci U S A* **102**, 13950-5 (2005).
7. Konstantinidis, K.T. & Tiedje, J.M. Genomic insights that advance the species definition for prokaryotes. *Proc Natl Acad Sci U S A* **102**, 2567-72 (2005).
8. Maharjan, R., Seeto, S., Notley-McRobb, L. & Ferenci, T. Clonal adaptive radiation in a constant environment. *Science* **313**, 514-7 (2006).
9. Atwood, K.C., Schneider, L.K. & Ryan, F.J. Periodic selection in Escherichia coli. *Proc Natl Acad Sci U S A* **37**, 146-55 (1951).
10. Atwood, K.C., Schneider, L.K. & Ryan, F.J. Selective mechanisms in bacteria. *Cold Spring Harb Symp Quant Biol* **16**, 345-55 (1951).
11. Feil, E.J. Small change: keeping pace with microevolution. *Nat Rev Microbiol* **2**, 483-95 (2004).
12. Hershberg, R. et al. High Functional Diversity in Mycobacterium tuberculosis Driven by Genetic Drift and Human Demography. *PLoS Biol* **6**, e311 (2008).
13. Cohan, F.M. & Koeppel, A.F. The origins of ecological diversity in prokaryotes. *Curr Biol* **18**, R1024-34 (2008).
14. Acinas, S.G. et al. Fine-scale phylogenetic architecture of a complex bacterial community. *Nature* **430**, 551-4 (2004).
15. Koeppel, A. et al. Identifying the fundamental units of bacterial diversity: a paradigm shift to incorporate ecology into bacterial systematics. *Proc Natl Acad Sci U S A* **105**, 2504-9 (2008).
16. Gevers, D. et al. Opinion: Re-evaluating prokaryotic species. *Nat Rev Microbiol* **3**, 733-9 (2005).
17. Kassen, R. & Rainey, P.B. The ecology and genetics of microbial diversity. *Annu Rev Microbiol* **58**, 207-31 (2004).
18. Wildschutte, H. & Lawrence, J.G. Differential Salmonella survival against communities of intestinal amoebae. *Microbiology* **153**, 1781-9 (2007).
19. Zubkov, M.V. & Tarran, G.A. High bacterivory by the smallest phytoplankton in the North Atlantic Ocean. *Nature* **455**, 224-6 (2008).
20. Suttle, C.A. Marine viruses--major players in the global ecosystem. *Nat Rev Microbiol* **5**, 801-12 (2007).
21. Thingstad, T. Elements of a theory for the mechanisms controlling abundance, diversity, and biogeochemical role of lytic bacterial viruses in aquatic ecosystems. *Limnol Oceanogr* **45**, 1320-1328 (2000).
22. Thingstad, T.F., Lignell, R. Theoretical models for the control of bacterial growth rate, abundance, diversity and carbon demand. *Aquat. Microb. Ecol.* **13**, 19-27 (1997).
23. Coleman, M.L. et al. Genomic islands and the ecology and evolution of Prochlorococcus. *Science* **311**, 1768-70 (2006).

24. Legault, B.A. et al. Environmental genomics of "Haloquadratum walsbyi" in a saltern crystallizer indicates a large pool of accessory genes in an otherwise coherent species. *BMC Genomics* **7**, 171 (2006).
25. Coleman, M.L. & Chisholm, S.W. Code and context: Prochlorococcus as a model for cross-scale biology. *Trends Microbiol* **15**, 398-407 (2007).
26. Wilhelm, L.J., Tripp, H.J., Givan, S.A., Smith, D.P. & Giovannoni, S.J. Natural variation in SAR11 marine bacterioplankton genomes inferred from metagenomic data. *Biol Direct* **2**, 27 (2007).
27. Cuadros-Orellana, S. et al. Genomic plasticity in prokaryotes: the case of the square haloarchaeon. *Isme J* **1**, 235-45 (2007).
28. Kettler, G.C. et al. Patterns and Implications of Gene Gain and Loss in the Evolution of Prochlorococcus. *PLoS Genet* **3**, e231 (2007).
29. Frias-Lopez, J. et al. Microbial community gene expression in ocean surface waters. *Proc Natl Acad Sci U S A* **105**, 3805-10 (2008).
30. Garcia Martin, H. et al. Metagenomic analysis of two enhanced biological phosphorus removal (EBPR) sludge communities. *Nat Biotechnol* **24**, 1263-9 (2006).
31. Sharma, R.S., Mishra, V., Mohmmed, A. & Babu, C.R. Phage specificity and lipopolysaccarides of stem- and root-nodulating bacteria (Azorhizobium caulinodans, Sinorhizobium spp., and Rhizobium spp.) of Sesbania spp. *Arch Microbiol* **189**, 411-8 (2008).
32. Fitzgerald, C., Sherwood, R., Gheesling, L.L., Brenner, F.W. & Fields, P.I. Molecular analysis of the rfb O antigen gene cluster of Salmonella enterica serogroup O:6,14 and development of a serogroup-specific PCR assay. *Appl Environ Microbiol* **69**, 6099-105 (2003).
33. Ivars-Martinez, E. et al. Comparative genomics of two ecotypes of the marine planktonic copiotroph Alteromonas macleodii suggests alternative lifestyles associated with different kinds of particulate organic matter. *Isme J* **2**, 1194-212 (2008).
34. Reva, O. & Tummler, B. Think big--giant genes in bacteria. *Environ Microbiol* **10**, 768-77 (2008).
35. Leiman, P.G. et al. The structures of bacteriophages K1E and K1-5 explain processive degradation of polysaccharide capsules and evolution of new host specificities. *J Mol Biol* **371**, 836-49 (2007).
36. Mizoguchi, K. et al. Coevolution of bacteriophage PP01 and Escherichia coli O157:H7 in continuous culture. *Appl Environ Microbiol* **69**, 170-6 (2003).
37. Wilmes, P., Simmons, S.L., Denef, V.J. & Banfield, J.F. The dynamic genetic repertoire of microbial communities. *FEMS Microbiol Rev* **33**, 109-32 (2009).
38. Sorek, R., Kunin, V. & Hugenholtz, P. CRISPR--a widespread system that provides acquired resistance against phages in bacteria and archaea. *Nat Rev Microbiol* **6**, 181-6 (2008).
39. Kunin, V. et al. A bacterial metapopulation adapts locally to phage predation despite global dispersal. *Genome Res* **18**, 293-7 (2008).
40. Cardinale, B.J. et al. Effects of biodiversity on the functioning of trophic groups and ecosystems. *Nature* **443**, 989-92 (2006).
41. Tylianakis, J.M., Didham, R.K., Bascompte, J. & Wardle, D.A. Global change and species interactions in terrestrial ecosystems. *Ecol Lett* **11**, 1351-63 (2008).
42. Thompson, J.R. et al. Genotypic diversity within a natural coastal bacterioplankton population. *Science* **307**, 1311-3 (2005).
43. Brown, M.V., Schwalbach, M.S., Hewson, I. & Fuhrman, J.A. Coupling 16S-ITS rDNA clone libraries and automated ribosomal intergenic spacer analysis to show marine microbial diversity: development and application to a time series. *Environ Microbiol* **7**, 1466-79 (2005).

44. Allen, E.E. et al. Genome dynamics in a natural archaeal population. *Proc Natl Acad Sci U S A* **104**, 1883-8 (2007).
45. Hunt, D.E. et al. Resource partitioning and sympatric differentiation among closely related bacterioplankton. *Science* **320**, 1081-5 (2008).
46. Hennes, K.P., Suttle, C.A. & Chan, A.M. Fluorescently Labeled Virus Probes Show that Natural Virus Populations Can Control the Structure of Marine Microbial Communities. *Appl Environ Microbiol* **61**, 3623-3627 (1995).
47. Wommack, K.E., Ravel, J., Hill, R.T. & Colwell, R.R. Hybridization analysis of chesapeake bay virioplankton. *Appl Environ Microbiol* **65**, 241-50 (1999).
48. Parada, V., Baudoux, A.C., Sintes, E., Weinbauer, M.G. & Herndl, G.J. Dynamics and diversity of newly produced virioplankton in the North Sea. *Isme J* **2**, 924-36 (2008).
49. Golomidova, A., Kulikov, E., Isaeva, A., Manykin, A. & Letarov, A. The diversity of coliphages and coliforms in horse feces reveals a complex pattern of ecological interactions. *Appl Environ Microbiol* **73**, 5975-81 (2007).
50. Holmfeldt, K., Middelboe, M., Nybroe, O. & Riemann, L. Large variabilities in host strain susceptibility and phage host range govern interactions between lytic marine phages and their Flavobacterium hosts. *Appl Environ Microbiol* **73**, 6730-9 (2007).
51. Shoresh, N., Hegreness, M. & Kishony, R. Evolution exacerbates the paradox of the plankton. *Proc Natl Acad Sci U S A* **105**, 12365-9 (2008).
52. Wittebolle, L. et al. Initial community evenness favours functionality under selective stress. *Nature* **458**, 623-6 (2009).
53. Benton, T.G., Solan, M., Travis, J.M. & Sait, S.M. Microcosm experiments can inform global ecological problems. *Trends Ecol Evol* **22**, 516-21 (2007).
54. Brockhurst, M.A., Fenton, A., Roulston, B. & Rainey, P.B. The impact of phages on interspecific competition in experimental populations of bacteria. *BMC Ecol* **6**, 19 (2006).
55. Maynard-Smith, J. Evolution and the theory of games. *University Press, Cambridge* (1982).
56. Simmons, S.L. et al. Population genomic analysis of strain variation in Leptospirillum group II bacteria involved in acid mine drainage formation. *PLoS Biol* **6**, e177 (2008).
57. Strom, S.L. Microbial ecology of ocean biogeochemistry: a community perspective. *Science* **320**, 1043-5 (2008).
58. Fraser, C., Hanage, W.P. & Spratt, B.G. Recombination and the nature of bacterial speciation. *Science* **315**, 476-80 (2007).
59. Schoenfeld, T. et al. Assembly of viral metagenomes from yellowstone hot springs. *Appl Environ Microbiol* **74**, 4164-74 (2008).
60. Rusch, D.B. et al. The Sorcerer II Global Ocean Sampling Expedition: Northwest Atlantic through Eastern Tropical Pacific. *PLoS Biol* **5**, e77 (2007).
61. Dinsdale, E.A. et al. Functional metagenomic profiling of nine biomes. *Nature* **452**, 629-32 (2008).
62. Qimron, U., Marintcheva, B., Tabor, S. & Richardson, C.C. Genomewide screens for Escherichia coli genes affecting growth of T7 bacteriophage. *Proc Natl Acad Sci U S A* **103**, 19039-44 (2006).
63. Forde, S.E., Thompson, J.N., Holt, R.D. & Bohannan, B.J. Coevolution drives temporal changes in fitness and diversity across environments in a bacteria-bacteriophage interaction. *Evolution* **62**, 1830-9 (2008).
64. Konstantinidis, K.T. & Tiedje, J.M. Towards a genome-based taxonomy for prokaryotes. *J Bacteriol* **187**, 6258-64 (2005).
65. Ochman, H., Lerat, E. & Daubin, V. Examining bacterial species under the specter of gene transfer and exchange. *Proc Natl Acad Sci U S A* **102 Suppl 1**, 6595-9 (2005).
66. Read, T.D. & Ussery, D.W. Opening the pan-genomics box. *Curr Opin Microbiol* **9**, 496-8 (2006).

67. Stent, G.S. Molecular biology of bacterial viruses. *W.H. Freeman & Co.* (1963).

Figure 1. Rodriguez-Valera *et al.*

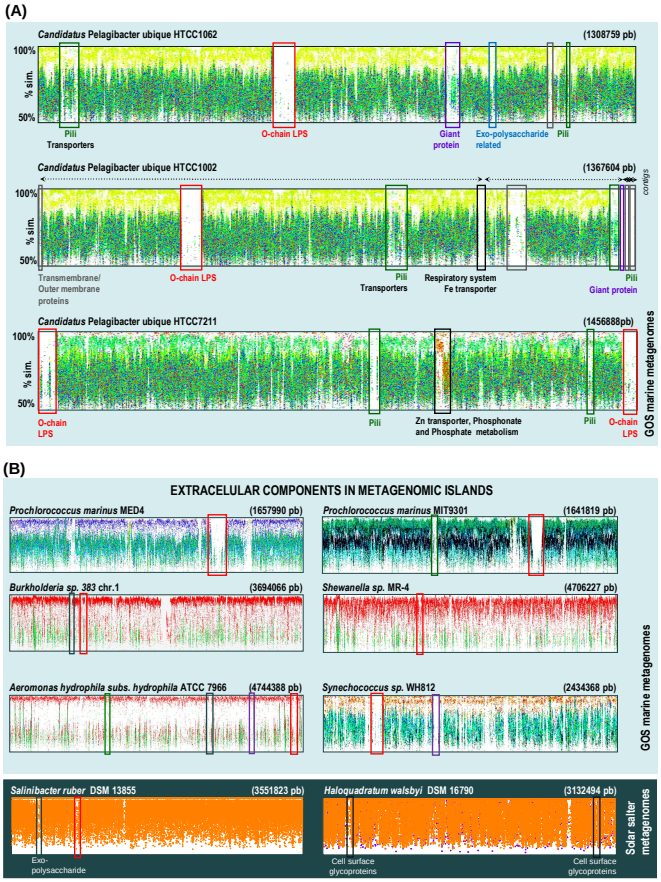


Figure 2. Rodriguez-Valera *et al.*

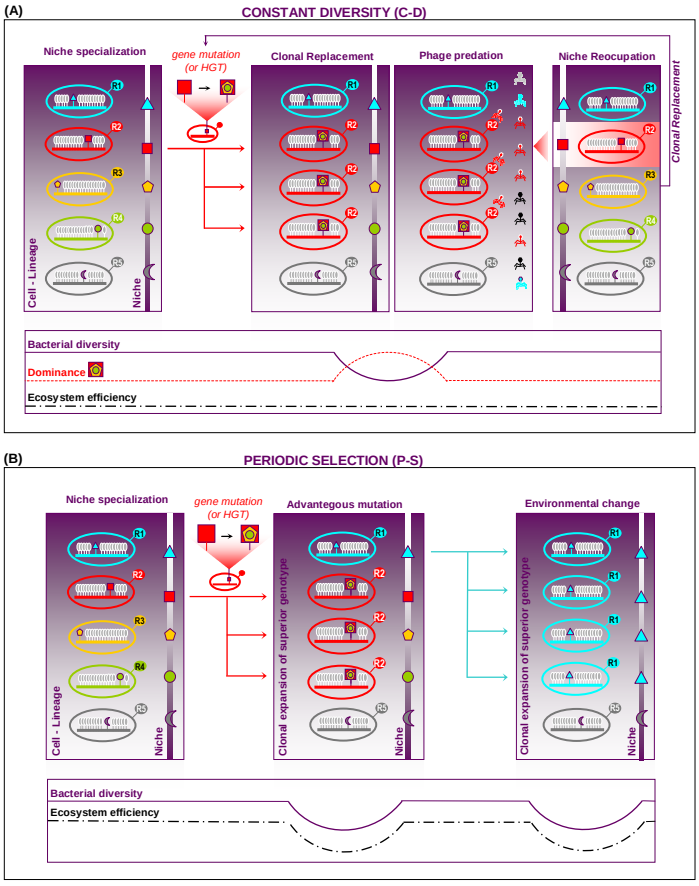


Figure 3. Rodriguez-Valera *et al.*