Trends in Microbiology

CellPress

Opinion

Mechanisms That Shape Microbial Pangenomes

Maria Rosa Domingo-Sananes [1,2,*] and James O. McInerney [1,*]

Analyses of multiple whole-genome sequences from the same species have revealed that differences in gene content can be substantial, particularly in prokaryotes. Such variation has led to the recognition of pangenomes, the complete set of genes present in a species – consisting of core genes, present in all individuals, and accessory genes whose presence is variable. Questions now arise about how pangenomes originate and evolve. We describe how gene content variation can arise as a result of the combination of several processes, including random drift, selection, gain/loss balance, and the influence of ecological and epistatic interactions. We believe that identifying the contributions of these processes to pangenomes will need novel theoretical approaches and empirical data.

Pangenomes and Why They Matter

The study of natural variation within and between species initially focused on phenotypes and later on genetic variation. Nucleotide sequence variation (NSV), in the form of SNPs and short indels, has been studied for decades, and has been analysed through the robust theoretical framework of population genetics, which aims to characterise and model genetic variation. High-throughput genome sequencing has made us aware of larger-scale variation between the genomes of the same species, and in particular the existence of extensive gene content variation (GCV), especially in prokaryotes [1–4]. Pangenomes, defined as the complete set of genes present in a species, encompass this diversity. Pangenomes contain core genes, that are present in all individuals, and accessory genes whose presence varies. The pangenome concept has been expanded to consider structural and copy-number variation in both protein-coding and noncoding sequences, particularly in eukaryotes [3,4]. Additionally, although pangenomes were originally conceived for a species, in principle we can apply the idea to any taxonomic unit, from a population to the pangenome of life [3]. In this article we focus on GCV in prokaryote species, although some of the mechanisms described here may also apply to eukaryotes and higher taxonomic levels.

Pangenomes arise as a consequence of constant gene gain and loss, the former commonly as a result of horizontal gene transfer (HGT) in prokaryotes [5–8]. These gains and losses are then subject to drift and selection, resulting in the typical patterns we observe in pangenomes (Figure 1). These patterns include an increase in the number of observed accessory genes and a decrease in the number of observed core genes as we sequence more genomes from the same species (Figure 1C) [1,9,10], as well as a U-shaped gene frequency distribution or spectrum (Figure 1D) [11,12]. However, the details of these patterns can vary considerably for different species. For example, as more genomes are sequenced the number of newly discovered genes can level off at very different points, and the proportion of core genes can vary significantly [11,13]. As a first approximation, these observations have led to pangenomes being classified as open or closed (Figure 1) [2,3,10], but metrics such as **genome fluidity** (see Glossary) (Figure 1E) have been proposed to better quantify pangenome diversity [11].

Highlights

The genomes of individuals of the same species can display large amounts of variation in gene content, particularly in prokaryotes. We still do not understand the reasons behind this diversity.

It is not clear to what extent the set of variable genes, the accessory genome, contributes to fitness. Different mechanisms can contribute to explain gene-content variation, including selection-dominated and random genetic drift-dominated processes.

Variability in rates of gene gain and loss and fitness likely plays an important role in explaining pangenome variability. The distribution of these parameters will likely vary for different species.

Multiple mechanisms likely contribute to gene content variation from neutral to selective, including gene gain/loss balance, gene-by-environment interactions, Black Queen dynamics and social interactions, and gene–gene interactions.

The mechanisms that contribute to gene content diversity likely vary within and between species and could be themselves subject to evolution and selection.

We are just starting to develop the theoretical toolkit required to describe and understand gene content variability and pangenomes.

Understanding gene-content variation and evolution is important to understand microbial adaptation and associated processes, such as emergence of antimicrobial resistance and new pathogens.

[1]School of Life Sciences, University of Nottingham, Nottingham, UK
[2]School of Science and Technology, Nottingham Trent University, Nottingham, UK

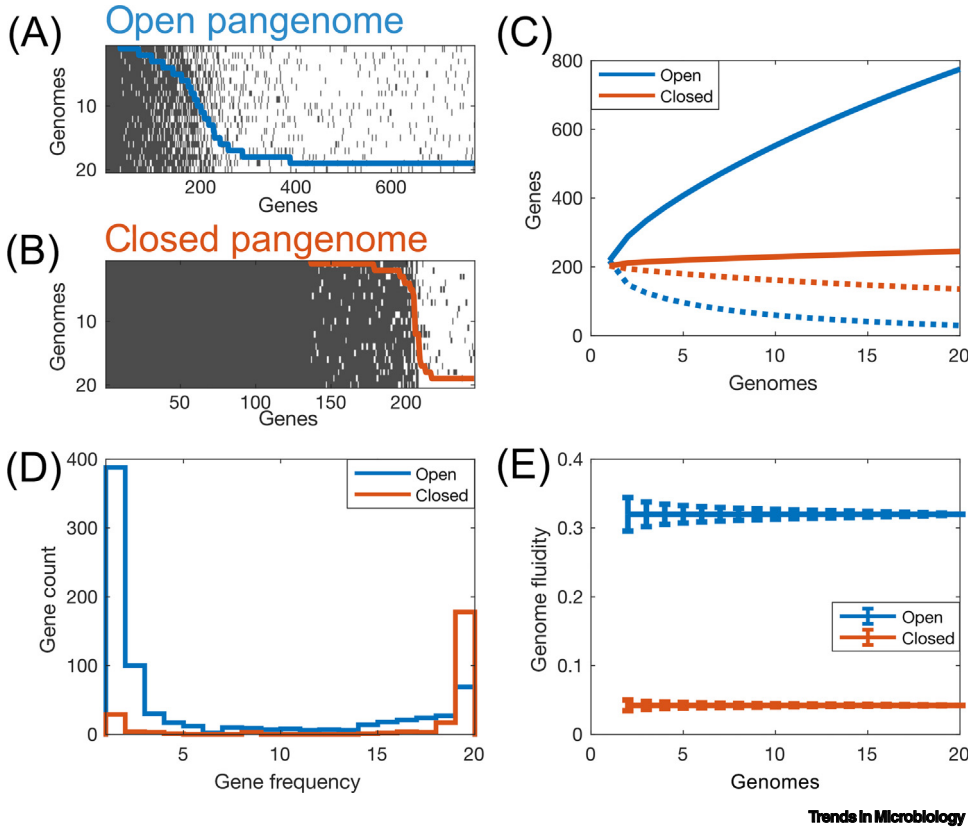

Figure 1. Properties of Open and Closed Pangenomes. Gene presence/absence (grey/white) for representations of an open (A) and a closed (B) pangenome, with genes sorted from most to least common. The blue and orange lines show the gene frequency. (C) Gene accumulation curves (unbroken lines) and core-gene depletion curves (broken lines) for the open (blue) and closed (orange) pangenomes from (A) and (B). (D) Gene frequency distributions for the open (blue) and closed (orange) pangenomes from (A) and (B). (E) Estimation of genome fluidity for the open (blue) and closed (orange) pangenomes from (A) and (B).

The significance of the variability in pangenome properties is still an open question. In particular, the extent to which accessory genes contribute to individual fitness is one of the most intriguing aspects of pangenomes, and the cause of recent debate [2,3,12,14,15]. Some accessory genes are likely to be genetic parasites, others neutral or nearly neutral, and some beneficial in at least some contexts [2,8,12,16,17]. An indication that genes of all these classes are present in pangenomes comes from the deletion of sets of accessory genes in *Escherichia coli* K-12 MG1655. Most deletions had neutral or deleterious effects on the bacterium's growth rate in rich media, indicating a neutral or beneficial role of these genes, although a few deleterious genes were also found [18,19]. Overall, however, we do not yet know the proportion of these different gene classes in pangenomes and their relationship with species-level gene frequency and species-level characteristics, such as overall prevalence of phage and mobile genetic elements, population size, and occupancy of different environmental niches. Understanding how GCV contributes to adaptation is not only interesting from an evolutionary perspective but it is important for predicting and understanding virulence, pathogenicity (infectiousness), and the spread of antimicrobial resistance. Additionally, GCV is important for understanding microbial ecology. For example, variable genes may have roles in adaptation to specific and changing environments, as seen for different ecotypes of the marine bacterium *Prochlorococcus*. In this diverse and highly abundant species, accessory genes are associated with specific conditions

*Correspondence:
maria.domingo-sananes@ntu.ac.uk (M. R. Domingo-Sananes) and
James.McInerney@nottingham.ac.uk (J.O. McInerney).

such as temperature, light, and phosphate and nitrogen availability [20,21]. Discovering genes that contribute to adaptation to different conditions could also be relevant for biotechnology.

As with sequence polymorphism, differences in gene content that lead to changes in microbial fitness can be acted on by natural selection, while the dynamics of (nearly) neutral variants can be explained by a combination of genetic drift and linkage with beneficial or deleterious mutations [3,8,12]. Drift can also dominate evolutionary dynamics in small populations. Both drift and directional selection are expected to continually remove variation, and it is therefore important to consider why we observe such extensive GCV. Variation can be partly explained by random evolutionary processes, in particular, clonal reproduction with gene gain and loss, as proposed by some neutral models [22,23] (Box 1). However, crucially, these neutral models often do not accurately fit real data, indicating that other mechanisms may have a role in shaping pangenomes [23–25] (Box 1). Here we propose and describe mechanisms that may contribute to generating and maintaining GCV, such as a balance between gene gain and loss and interactions between accessory genes and ecological/genetic factors.

Parameters That Shape Pangenomes

To understand why accessory genes exist, what determines their frequency, and why we see the patterns presented in Figure 1, we need to consider several processes and parameters. As with NSV, the simplest null model we can consider is one in which all GCV is neutral. In this case we expect that populations with a larger **effective population size** (N_e) will manifest a greater amount of variability [12,26,27]. Genome fluidity and pangenome size are both correlated with neutral sequence variation [26,28], which is, in turn, a proxy of effective population size, and

Box 1. Neutral Models of Pangenome Evolution

The first step towards the development of a theoretical framework for the evolution of GCV and pangenomes is an appropriate neutral model. Two main approaches have been developed. Haegeman and Weitz [22] developed an individual-based model in which a population of cells is simulated by a birth-and-death process, while gaining and losing genes. In this model, the genome size is fixed (a lost gene is replaced by a new one from the environment), each acquired gene is new to the population, and the rates of gain and loss are the same for all genes and cells. This simple model can recover U-shaped gene frequency distributions and the typical shapes of gene-accumulation and core-gene depletion curves. However, it does not entirely capture these features when compared with real pangenomes from multiple species. In particular, the model tends to predict fewer rare genes and more common genes than observed in real pangenomes. Adding two classes of genes with different loss rates helps to improve the fit [22]. Since genome size does vary within species, and gene transfer can occur within a population, incorporating these features might improve the explanatory power of this model.

Another approach is the infinitely many genes model [23], named after the infinitely many alleles model of sequence evolution [69]. As in the previous case, the original formulation of the model assumes that every gene can be acquired only once. Gene gains and losses are modelled along a phylogenetic tree. This tree can be a random tree (simulated based on population parameters), or a tree inferred from sequence data. This model also recovers the general expected shapes of gene frequency distributions and gene-accumulation and core-gene depletion curves. However, again, the exact patterns of real data do not fit well to this model. The fit improves substantially when gains and losses are modelled along an inferred phylogeny rather than a random tree [23]. Further improvement can be achieved by explicitly incorporating core genes (genes that cannot be gained or lost) [24]. This reflects the importance of selection in maintaining core, essential genes, which is perhaps unsurprising. The infinitely many genes model with two gene classes – highly mobile and immobile (essential) genes – fits some real data very well, such as the gene accumulation curves for *S. pneumoniae*, that are representative of the number of additional genes found as the number of analysed genomes increases [24]. However, the fit is less accurate for the core-gene depletion curve, where the model overestimates the number of core genes [24]. Interestingly, adding a third class of genes with intermediate rates of gene gain and loss (and thus mobility) improves the fit even further [24]. These gene classes of high, medium, and no mobility might correspond respectively to the proposed classes of accessory genes based on their frequency in the pangenome: cloud (singletons or very low frequency genes), shell (intermediate frequency genes, e.g., 10–99%), and core genes (present in all individuals) [7,24]. The proposed intermediate motility of shell genes suggests that they may be more likely to be maintained at intermediate frequencies in the population through selection and the other mechanisms described in the main text.

Glossary

Black Queen hypothesis: proposes that loss of genes encoding useful 'leaky' functions (usually production of public goods) can occur if other members of the community can provide such function. Multiple losses of this type within a community or population can lead to dependencies between organisms to complete functional processes, which thus become partially encoded in different cells [48]. This mechanism could contribute to the maintenance of genes at stable intermediate frequencies.
Effective population size (N_e): the size of an idealised population that has the same amount of genetic variation or experiences the same amount of genetic drift as an observed population. N_e is usually much smaller than the census or real population size [70].
Gene gain/loss balance: a mechanism that could maintain genes at stable intermediate frequencies in a group of organisms when the rates of gene gain and/or loss are high.
Gene-by-environment interactions: describes how the fitness effect of a genetic variant (gene) can depend on the environment. For example, a variant can be beneficial in one environment and deleterious in another. Across a group of organisms living in different environments, these interactions could maintain genes at stable intermediate frequencies.
Gene–gene interactions: a class of epistatic interactions in which the fitness effects of genes are dependent on other genes – for example, if two genes are deleterious when present in isolation but beneficial when present together; this type of interaction could result in increased pangenome diversity.
Genome fluidity: a measure of the distance or dissimilarity in gene content between genomes. For a pair of genomes, it is the ratio between the number of genes that are not shared between them and the total number of genes in both genomes. For a pangenome, genome fluidity is the average of the pairwise measures [11].
Negative frequency-dependent selection (NFDS): occurs when the fitness effect of a variant decreases as its frequency in the population increases. This mechanism can maintain genes at stable intermediate frequencies within a population.
Positive frequency-dependent selection: occurs when the fitness contribution of a variant increases as it

this has been taken as evidence that a large proportion of GCV might be neutral [28]. Additionally, mathematical models of neutral gene content evolution [22,23] are able to recover, to an extent, some of the patterns observed in Figure 1, but these models may not completely account for the extent of GCV observed, at least in some species [24,25] (Box 1).

However, there is growing empirical evidence that shows that many accessory gene changes are not neutral with respect to the fitness of the host cell [29–31]. Every gene is associated with a particular 'fitness effect', or contribution towards its host. For the many genes that can be acquired, there is an associated 'distribution' of fitness effects (DFE) [8,12]. Though we know that such a distribution exists, we do not have precise measurements for what these DFEs look like for both incoming genes and for the cohorts of genes that are lost. The shapes of these distributions could be similar to the DFEs of mutations, but they could also be very different (Figure 2A,B and Box 2). Furthermore, in species with large N_e, selection is expected to be highly effective and contribute more to reducing the frequency of slightly deleterious genes and increasing that of slightly beneficial genes [2,26,27]. This combination of theoretical insights and empirical evidence has led to the proposal that the correlation between N_e and pangenome size is due to selection [2,26]. In this case, most accessory genes would be expected to be beneficial because slightly beneficial genes are more likely to be maintained by selection in large populations [26] and because species with large N_e tend to occupy a wider range of ecological niches where different sets of accessory genes may be selected for [32].

In terms of how genetic changes arise, there is a fundamental difference between the processes that generate NSV (mutation) and GCV (gene gain and loss). While mutation rates are roughly similar for any given genome, different genes could be 'physically' gained and lost by individual genomes at vastly different rates. This is because different gene gain/loss mechanisms occur with different frequencies. For example, a gene associated with a transposon, or located in a conjugative plasmid, has a higher potential for transfer than a gene not associated with a mobile genetic element. An alarming example is the *mcr-1* gene encoding colistin resistance, whose association with a transposon likely enabled its rapid spread across the world and its occurrence in multiple species [31]. Genes associated with plasmids and transposons would likely also have higher rates of loss. For individual genes, gain/loss rates can have a dramatic influence on their frequency in a population. For example, a gene that is gained at high rates can be acquired multiple times, resulting in its spread and maintenance in a population, even if it is deleterious. These kinds of genes are typically known as selfish genetic elements (Box 2) [16,33–35]. On top of variation for individual genes, different species can vary in the intrinsic rates of gene gain and loss. For example, naturally competent bacteria likely have higher gain rates, which may correlate with larger pangenome size [6,36], while the rates of homologous recombination are known to vary in different species [37], potentially leading to differences in the rates of loss. However, the presence of restriction–modification systems, or CRISPR systems, can mitigate particular kinds of gene gains, even before drift or selection has any effect [37,38].

Overall, because rates of gain/loss vary for different genes, for the pool of genes that can be gained and lost, 'distributions' of these parameters exist. Furthermore, the shape of these distributions likely varies in different organisms, which, in turn, may lead to differences in the properties of their pangenomes. As with the DFE, we know little about what these distributions of gain and loss for different genes may look like. In most theoretical frameworks of pangenome evolution, rates of gain and loss are assumed to be the same for all genes, or for sets of two or three gene classes [22,23,39]. However, we can consider different possible distributions using a simple model of pangenome evolution (Box 2) which predicts that more variable rates of gene gain and loss should result in more variable pangenomes (Figure 2C,D) [34]. Knowing more about real distributions of gene gain and loss would let us assess how much they influence pangenome properties.

becomes more common in the population. This mechanism can cause fast divergence between populations of the same species, and thus contribute to increased variation.

Second-order selection: natural selection acting on the parameters that can determine the rate of evolution and adaptation, such as the rates of mutation, recombination, and gene gain and loss [71].

Social interactions: interactions between organisms of the same or different species that can affect their fitness. Interactions can be mutually beneficial, altruistic, selfish, or spiteful [72].

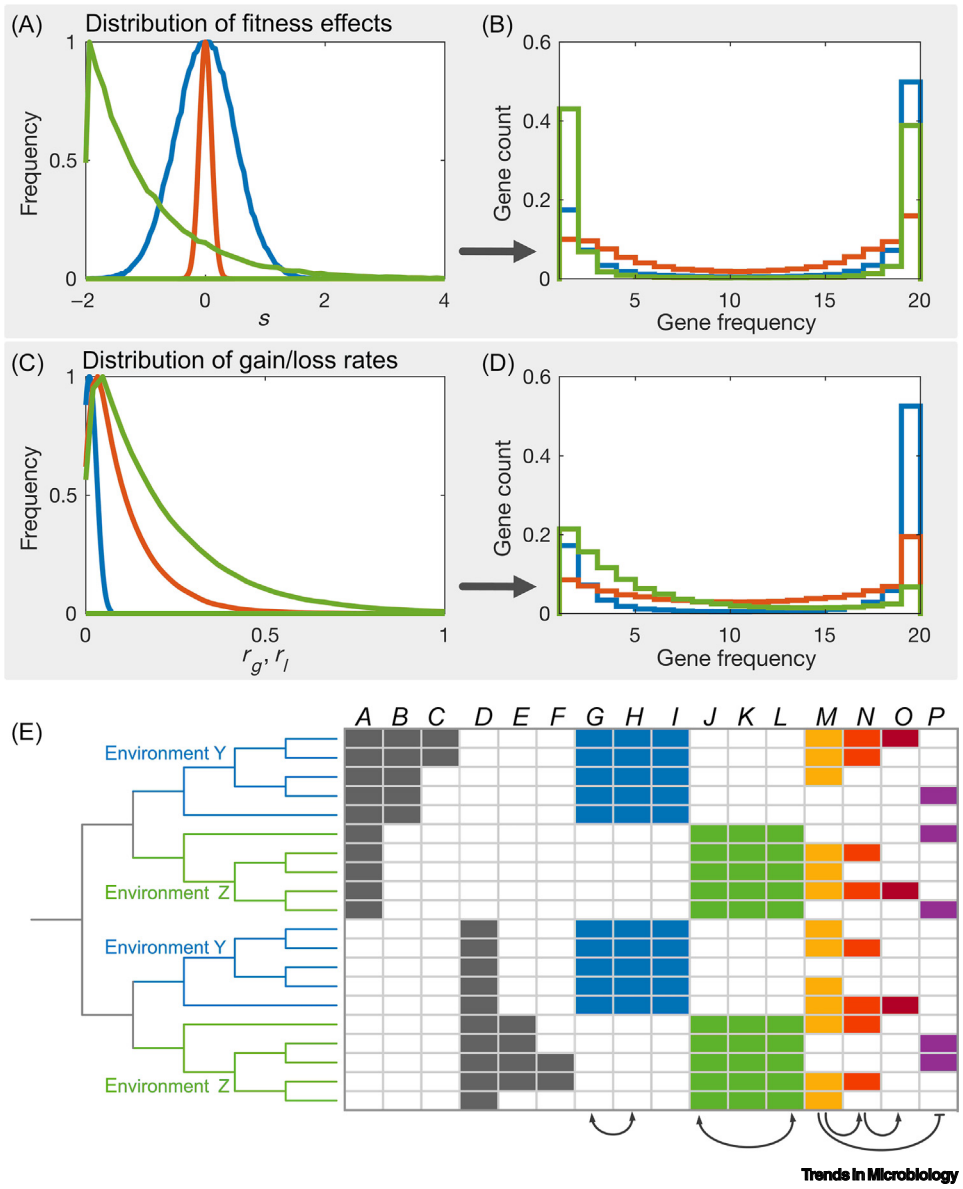

Figure 2. Parameters and Interactions That Shape Pangenomes. (A) Examples of possible types of distributions of fitness effects of genes that can be gained or lost, and (B) corresponding expectations for the gene frequency distribution (from the model described in Box 1). (C) Examples of possible types of distributions of rates of gene gain and loss, and (D) corresponding expectations for the gene frequency distribution (from the model described in Box 1). (E) Schematic examples of interactions that contribute to gene content variation: grey genes (*A–F*) are dependent on phylogeny; blue (*G–I*) and green (*J–L*) genes are associated with the environments from which the genomes were sampled, although they may also interact directly with each other (black arrows at the bottom); for instance, the presence of gene *O* (red) is conditional on the presence of gene *N* (orange), which is conditional on the presence of gene *M* (yellow); the presence of gene *P* (purple) is conditional on the absence of gene *M* (yellow).

Importantly, high rates of gene gain and loss imply that some genes could be acquired and/or lost multiple times in different genetic backgrounds. This means that a balance between gain and loss could maintain some genes at stable intermediate frequencies in populations and species (Box 2) [16,34,35,40]. Maintenance of stable polymorphisms by **gene gain/loss balance** may be much

Box 2. A Toy Model to Understand How Parameters May Affect Pangenomes

In order to assess the contributions of fitness effects of genes along with variability in their gain/loss rates, we can consider a simple mathematical model. For a single gene that can be gained and lost we assume an additive contribution to fitness, s, which can be positive or negative, and gene-specific rates of gain, r_g and loss r_l. Then the frequency of the gene in a population of cells can be described by a differential equation [34] (and an approach similar to that found in [35]):

$$\frac{dx}{dt} = r_g(1-x) + sx(1-x) - r_l x \quad [I]$$

From this, we can plot the steady state of gene frequency with respect to fitness contribution for different values of gain/loss rates (Figure I). In general, beneficial genes would be expected to be found at higher frequencies, while deleterious genes would be present at lower frequencies. But, as described in the main text, if the rates of gain and loss are high, genes can be maintained at intermediate frequencies (gain/loss balance), and specifically, deleterious genes may be found at relatively high frequencies, while even highly beneficial genes may not be fixed in the population. Assuming no interactions between genes, we can use this model to test the effect that different distributions of fitness effects and rates of gene gain and loss may have on pangenomes (see Figure 2A–D in main text; [34]). Although this simple model can give us some insight, it does not capture the contribution of the evolutionary process described by the models presented in Box 1. In particular, the model considers genes to be independently gained and lost, and therefore does not consider genome-wide linkage, in contrast to the models described in Box 1. Future theoretical analyses should aim to bridge the gap between these approaches in order to develop a comprehensive theoretical framework for pangenomes and their evolution.

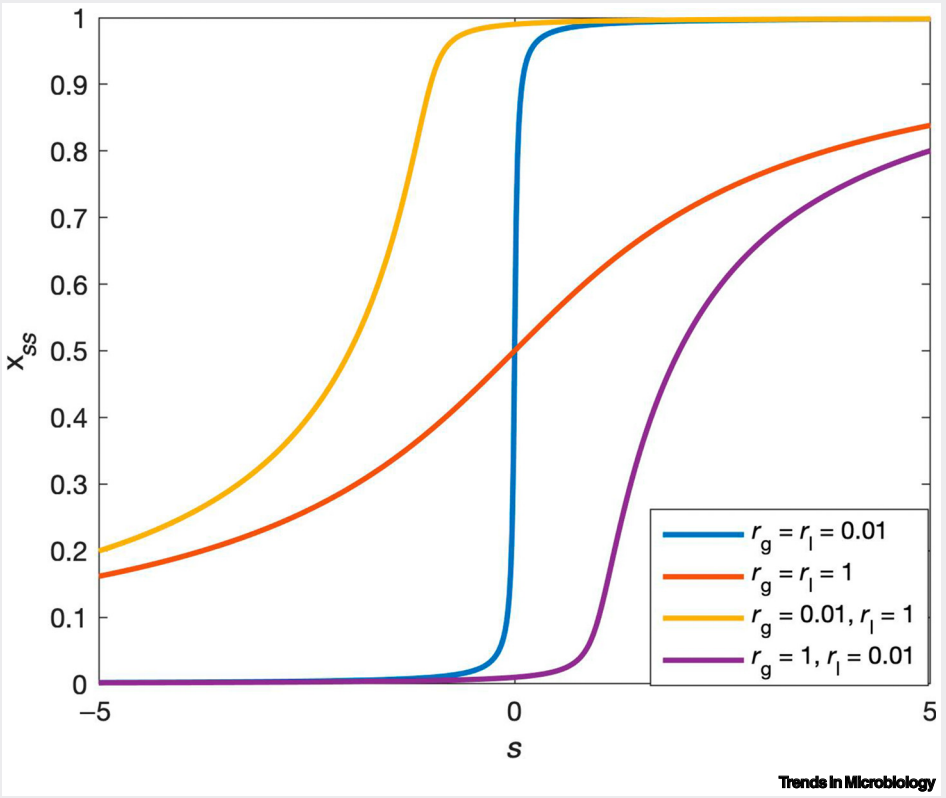

Figure I. Equilibrium Gene Frequency with Respect to Fitness Effect According to the Model Described in Box 2. The x-axis represents the contribution of a gene to the fitness of the cell, while the y-axis indicates the expected frequency of the gene in a population. The lines indicate the gene frequency that would be expected exclusively under gain/loss balance, that is, the steady state described by the model. Different lines show different combinations of rates of gene gain and loss.

more important in GCV compared with NSV. This is because occurrence and eventual maintenance of the same mutation multiple times in different backgrounds is probably a rare event, and multiple reversion mutations (the equivalent of gene loss) should be very unlikely. A drawback of this constant gain and loss and variability in the rates of these events is that we cannot directly link the fitness effect of a gene to its frequency or dynamics in the population. Gain/loss balance is therefore a mechanism that should be taken into account when modelling and analysing pangenomes [41], along with variability in gene fitness and gain/loss rates.

Interactions That Shape Pangenomes

So far, we have considered the simplified view of genes as independent entities associated with their own fitness effects, and rates of gain and loss. However, interactions between accessory genes and ecological factors can also affect GCV. **Gene-by-environment interactions** occur when a gene is beneficial in one environment but deleterious or neutral in others, a situation that is often seen for antibiotic-resistance genes [42]. For example, an unstable plasmid encoding a kanamycin/neomycin resistance gene in *Pseudomonas aeruginosa* is costly for cells and is rapidly lost in the absence of antibiotic, while the plasmid becomes beneficial and maintained in the presence of antibiotic [43]. Maintenance of the plasmid due to selection for antibiotic resistance then allows compensatory evolution to reduce the cost of carrying the plasmid, showing that the fitness effect of a particular gene can vary across time, even when the external ecosystem remains constant [43]. Due to these gene-by-environment interactions, some accessory genes may only be acquired and maintained in specific ecosystems or under certain conditions. This ecosystem-specific selective pressure can result in a gene becoming fixed or close to fixation, that is, reaching a frequency close to 1, but only in that ecosystem. Therefore, across the larger population, or at the species level, these genes could be present at low frequency, and consequently, they are considered to be accessory genes in the pangenome (Figure 2E, genes *G–L*; Box 2) [44]. If there is constant migration between ecological niches, these niche-specific genes may be acquired multiple times, potentially in different strain backgrounds, resulting in gene frequencies being at intermediate levels due to gain/loss balance [34,35,40] as described in the previous section. An illustrative example of this interaction between genes and the environment was recently shown in the yeast *Saccharomyces cerevisiae*, where introduction of a gene encoding a glycerol transporter that was transferred between different fungal clades conferred a fitness benefit to cells growing in high glycerol concentrations, but was deleterious for cells growing in glucose [44]. Another key example involves the bacterium *Campylobacter jejuni* in which the presence of a seven-gene region is associated with host preference and not phylogeny. Three of the genes in this region are involved in vitamin B_5 biosynthesis, which can, in turn, be beneficial to cattle, which have diets that are poor in vitamin B_5 [29]. As more data accumulates, it will be interesting to quantify what proportion of GCV is shaped by these environment- or niche-dependent effects.

Interactions between organisms can also influence GCV and contribute to the accessory genome, as outlined below. A classic case is **negative frequency-dependent selection** (**NFDS**), where a genetic variant is beneficial when it is relatively rare or below a certain frequency [45]. A hypothetical example is a gene encoding a surface protein that is beneficial to the microorganism, such as a nutrient transporter, but that is also a receptor for a phage. If the gene is present in most genomes, the population will be susceptible to the phage, resulting in lower absolute fitness, but if the gene is rare, the phage will not be able to spread and consequently, those cells that carry that particular gene would reap its benefit [45]. In this way, NFDS results in genes being stably maintained at intermediate frequencies. Analysis of the dynamics of pangenomes of *Streptococcus pneumoniae* [46] and different *E. coli* sequence types [47] suggests that NFDS maintains some accessory genes at stable intermediate frequencies, even

if the genetic backgrounds in which these genes are present change. However, it is possible that some of the other mechanisms described here, such as gain/loss balance, may also play a role.

Social interactions may also contribute to GCV. Genes encoding public goods may be subject to **positive frequency-dependent selection**, enabling rapid divergence between populations. A further mechanism that has been proposed to contribute to GCV is the distributed genome hypothesis or **Black Queen hypothesis** [48,49]. The idea behind this hypothesis is that 'leaky' functions, such as the production of a useful but excreted metabolite, or other public goods, can be lost in some cells if the rest of the population or community can provide the same function (that is associated with these goods/metabolites). This may have the benefit of allowing organisms to maintain a smaller genome [50], as has been proposed for some oligotrophic marine bacteria, such as *Prochlorococcus* and *Candidatus* Pelagibacter ubique [48]. Furthermore, interactions between members of the community that perform different functions could lead to stable populations where multiple genes are maintained at intermediate frequencies [48,49]. While, for many bacteria, a reduction in the size of the genome may not be of direct benefit [19,39,51,52], additional factors could encourage interactions similar to those proposed by the Black Queen hypothesis, such as compartmentalising functions in cells with the most appropriate genetic backgrounds or allowing division of labour. While these types of complex social interaction may be rare, they have been detected, for example, as metabolic cross-feeding, where individuals exchange metabolites that benefit one or both partners [53]. Most instances of cross-feeding are observed between species, but they may occur within populations of the same species [53].

Gene–gene interactions within a genome may also contribute to GCV and the complex patterns that we observe within pangenomes [54–56] (Figure 2E). The simplest case involves a pair of genes that have different fitness contributions when both are found together, compared with when each gene is present on its own (that is, nonadditive contributions to fitness). The fitness effect of being jointly present might be an overall positive or a negative effect. These types of interaction are not confined to gene pairs and could occur across groups of genes. Conditional relationships, where the gain or maintenance of a particular gene is more likely when another gene is present, may also occur within genomes (Figure 2E, genes *M–P*). Furthermore, these relationships could also occur between accessory genes and particular sequence variants (or combinations of variants). Such sequence variants could be present in core genes, or in other accessory genes. Associations of this type were recently observed in the pathogen *Vibrio parahaemolyticus*. It was suggested that these associations could contribute to distinct ecological strategies in the marine environments that the bacterium inhabits [56]. These intricate epistatic interactions could lead to complex patterns of gene presence, including relatively stable intermediate frequencies, along with co-occurrence, avoidance, and dependency relationships between genes in pangenomes. Analysis of multiple genomes and pangenomes have confirmed the existence of these patterns [54–58], although the prevalence of these interactions and their influence on phenotypes, fitness, and evolution are not yet clear.

Although patterns of gene co-occurrence or avoidance may occur and be relatively common, we should be cautious of ascribing them to direct gene–gene interactions since these patterns can also arise as a consequence of external environmental influences. That is, natural selection could cause two or more genes to co-occur if they are advantageous in the same environment, even if their functions in the cell are unrelated. Similarly, gene avoidance could result from two genes that do not interact, being simply unable to operate in a particular ecosystem (Figure 2E, genes *G–L*), while nested environments or niches could lead to nested gene sets (similar to gene dependencies) for genes that do not have functional relationships. In addition, different

types of interaction may occur together. Further analysis should focus on dissecting how common all these types of interaction are and how significantly they affect pangenome properties and evolution.

The Evolvability of Pangenomes

As discussed in the first section, rates of gene gain and loss may vary for different genes, with some genes capable of promoting their own acquisition (e.g., transposons). Additionally, we can see variations in the uptake of foreign DNA across different species and across individuals within a specific species. For example, the distribution of competence across bacteria seems to be patchy [59], while the efficiency of transformation can vary among different strains of the same species, as shown in *S. pneumoniae* [60]. At least a proportion of the variation in rates of gene gain by competence is therefore genetically determined by the cell. Other host-encoded mechanisms controlling gene gain include the presence of defence or repair mechanisms such as restriction–modification systems or phage defence mechanisms such as CRISPR [37,38]. Additionally, gene loss is known to vary between species [61–63]. These observations suggest that there is variation in overall rates of gene gain and loss between different species and individuals of the same species. Theoretically, if there is variation in the rates of gene gain and loss between individuals, it could be acted upon by selection [8].

Species with higher rates of gene gain and large, open pangenomes may be able to occupy more niches [2,3] and be more adaptable. Indeed, species with higher genome fluidity seem to occupy a wider range of environmental niches [13]. However, this correlation may also be a consequence of large population sizes [26,28]. In an attempt to draw a parallel between GCV and NSV, in terms of comparing the rates of events that generate variation – mutation and gene gain/loss – we know that, at least under some conditions, elevated mutation rates can be selected for due to the increase in the supply of beneficial mutations [64]. However, elevated mutation rates can also result in the accumulation of many other neutral and slightly deleterious mutations [65]. The potential benefit of elevated mutation rates also depends on relatively low recombination since the mutation that caused the higher mutation rate in the first place must remain linked to the beneficial mutation(s) [65]. A similar benefit could be observed for elevated gene gain rates in nature. In novel environments, acquiring niche-specific genes may be highly beneficial but it can also come at the cost of acquiring deleterious or infectious genes.

Rates of gene loss are relatively high in prokaryotes and also vary between species [52,61,63]. The most intuitive explanation for high loss rates is that maintaining genes that do not provide a fitness benefit is costly, and therefore individuals that lose these genes will have a fitness advantage [50]. However, another possibility is that losing genes is not beneficial in itself but that loss rates are high, and as a consequence, the genes that remain in the genome are the most beneficial ones (because strains that lose those beneficial genes are at a disadvantage) [39,51,63,66]. The extent of this loss bias varies between organisms [52,63], which begs the question of what determines loss rates and why they are high. Are high loss rates maintained by selection? Another possible explanation for high gene loss rates is the existence of a 'drift barrier' similar to that proposed for the evolution of mutation rates [67]. In the case of pangenomes, this barrier means that slightly beneficial genes may be lost through genetic drift [26]. It also means that genetic drift may prevent the evolution of mechanisms (such as better DNA repair) to prevent such losses.

We still know very little about **second-order selection** on pangenomes, that is, selection operating on the rates of gene gain and loss. As demonstrated by the occurrence of mutator strains, for example, in *E. coli* adapting to a new host [68], recombination in prokaryotes may be sufficiently low for second-order selection [64,65]. In addition, gene gain and loss and gene content variation

may be more prone to second-order selection than NSV, partly because of the diversity of mechanisms responsible for DNA acquisition and deletion. If selection on the rates of gene gain and loss can be demonstrated, it will be interesting to see at which time scales it takes place, and how important it is for pangenome diversity and evolution.

Concluding Remarks and Future Perspectives

We still do not know what proportion of accessory genes are beneficial for the carrier cells in their particular environment. We do not know how stable pangenomes are: what proportion of variable genes are permanently polymorphic, and how many are in the process of being fixed or lost? There are several different explanations for GCV and intermediate gene frequencies. For example, mostly clonal evolution of large populations, combined with constant gain and loss of neutral genes, might be enough to explain the diversity of some pangenomes. However, stable intermediate frequencies may be maintained for some accessory genes due to high gain/loss rates (gain/loss balance), niche dependence (gene-by-environment interactions), interactions with other members of the population/other organisms (frequency-dependent selection, Black Queen dynamics), or epistasis. Furthermore, many different combinations of all these mechanisms are also possible, and their effects likely vary in different groups of organisms. A major question is whether we can quantify the importance of these diverse mechanisms, that is, what proportion of variation can be allocated to different processes and interactions [12]? Future work should identify which mechanism(s) best explain the presence or absence of individual genes, knowledge that could have important implications for medicine, ecology, and biotechnology (see Outstanding Questions). To accomplish these goals, we need to develop a testable theoretical framework that can capture the processes and mechanisms that we have considered. One possibility is to use modelling approaches based on the 'infinitely many genes' model [23] to test the effects of variation on gene fitness and gain/loss rates during pangenome evolution, as well as the consequences of different proportions of genes affected by the mechanisms described here. In order to be able to test these models and define the main contributors of GCV, we also need to acquire and analyse more whole genomes with associated metadata (e.g., phenotypic characteristics and properties of the environments where strains are isolated). Direct observation of pangenome evolution from longitudinal end experimental studies will also help us to disentangle the mechanisms that shape microbial pangenomes.

Outstanding Questions

How stable are pangenomes and their properties?

Are some genes maintained by selection at stable intermediate frequencies?

What proportion of accessory genes are neutral, adaptive (beneficial), or deleterious (genetic parasites)? Does this proportion vary between species?

Is effective population size the only major determinant of genome fluidity?

What are the shapes of the distribution of fitness effects and of rates of gene gain and loss?

How structured are populations, and how prevalent are gene-by-environment effects?

Does selection act on rates of gene gain and loss? If so, are optima variable and dependent on lifestyle, environment and/or taxonomic properties?

What determines loss rates and genome sizes? Does selection play a role?

Acknowledgments

Thanks to the members of the McInerney group for discussions, and to Elizabeth Cummins, Rebecca J. Hall, Fiona J. Whelan, and two anonymous reviewers for comments on the manuscript. This work was supported by the Biotechnology and Biological Sciences Research Council (grant BB/N018044/1).

References

1. Vernikos, G. *et al.* (2015) Ten years of pan-genome analyses. *Curr. Opin. Microbiol.* 23, 148–154
2. McInerney, J.O. *et al.* (2017) Why prokaryotes have pangenomes. *Nat. Microbiol.* 2, 1–5
3. Brockhurst, M.A. *et al.* (2019) The ecology and evolution of pangenomes. *Curr. Biol.* 29, R1094–R1103
4. Sibbald, S.J. *et al.* (2020) Lateral gene transfer mechanisms and pan-genomes in eukaryotes. *Trends Parasitol.* 36, 927–941
5. Treangen, T.J. and Rocha, E.P.C. (2011) Horizontal transfer, not duplication, drives the expansion of protein families in prokaryotes. *PLoS Genet.* 7, e1001284
6. Puigbò, P. *et al.* (2014) Genomes in turmoil: quantification of genome dynamics in prokaryote supergenomes. *BMC Biol.* 12, 66
7. Koonin, E.V. and Wolf, Y.I. (2008) Genomics of bacteria and archaea: the emerging dynamic view of the prokaryotic world. *Nucleic Acids Res.* 36, 6688–6719
8. Vos, M. *et al.* (2015) Rates of lateral gene transfer in prokaryotes: high but why? *Trends Microbiol.* 23, 598–605
9. Tettelin, H. *et al.* (2005) Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: Implications for the microbial 'pan-genome'. *Proc. Natl. Acad. Sci. U. S. A.* 102, 13950–13955
10. Medini, D. *et al.* (2005) The microbial pan-genome. *Curr. Opin. Genet. Dev.* 15, 589–594
11. Kislyuk, A.O. *et al.* (2011) Genomic fluidity: an integrative view of gene diversity within microbial populations. *BMC Genomics* 12, 32
12. Rocha, E.P.C. (2018) Neutral theory, microbial practice: challenges in bacterial population genetics. *Mol. Biol. Evol.* 35, 1338–1347
13. Maistrenko, O.M. *et al.* (2020) Disentangling the impact of environmental and phylogenetic constraints on prokaryotic within-species diversity. *ISME J.* 14, 1247–1259
14. Vos, M. and Eyre-Walker, A. (2017) Are pangenomes adaptive or not? *Nat. Microbiol.* 2, 1576
15. Shapiro, B.J. (2017) The population genetics of pangenomes. *Nat. Microbiol.* 2, 1574
16. Iranzo, J. *et al.* (2016) Inevitability of genetic parasites. *Genome Biol. Evol.* 8, 2856–2869
17. Nakamura, Y. *et al.* (2004) Biased biological functions of horizontally transferred genes in prokaryotic genomes. *Nat. Genet.* 36, 760–766

18. Pósfai, G. *et al.* (2006) Emergent properties of reduced-genome *Escherichia coli*. *Science* 312, 1044–1046
19. Karcagi, I. *et al.* (2016) Indispensability of horizontally transferred genes and its impact on bacterial genome streamlining. *Mol. Biol. Evol.* 33, 1257–1269
20. Coleman, M.L. *et al.* (2006) Genomic islands and the ecology and evolution of Prochlorococcus. *Science* 311, 1768–1770
21. Kent, A.G. *et al.* (2016) Global biogeography of *Prochlorococcus* genome diversity in the surface ocean. *ISME J.* 10, 1856–1865
22. Haegeman, B. and Weitz, J.S. (2012) A neutral theory of genome evolution and the frequency distribution of genes. *BMC Genomics* 13, 196
23. Baumdicker, F. *et al.* (2012) The infinitely many genes model for the distributed genome of bacteria. *Genome Biol. Evol.* 4, 443–456
24. Collins, R.E. and Higgs, P.G. (2012) Testing the infinitely many genes model for the evolution of the bacterial core genome and pangenome. *Mol. Biol. Evol.* 29, 3413–3425
25. Lobkovsky, A.E. *et al.* (2013) Gene frequency distributions reject a neutral model of genome evolution. *Genome Biol. Evol.* 5, 233–242
26. Bobay, L.-M. and Ochman, H. (2018) Factors driving effective population size and pan-genome evolution in bacteria. *BMC Evol. Biol.* 18, 153
27. Charlesworth, B. (2009) Effective population size and patterns of molecular evolution and variation. *Nat. Rev. Genet.* 10, 195–205
28. Andreani, N.A. *et al.* (2017) Prokaryote genome fluidity is dependent on effective population size. *ISME J.* 11, 1719–1721
29. Sheppard, S.K. *et al.* (2013) Genome-wide association study identifies vitamin B5 biosynthesis as a host specificity factor in *Campylobacter*. *Proc. Natl. Acad. Sci. U. S. A.* 110, 11923–11927
30. Lee, M.C. and Marx, C.J. (2012) Repeated, selection-driven genome reduction of accessory genes in experimental populations. *PLoS Genet.* 8, e1002651
31. Wang, R. *et al.* (2018) The global distribution and spread of the mobilized colistin resistance gene mcr-1. *Nat. Commun.* 9, 1179
32. Mclnerney, J.O. *et al.* (2020) Pangenomes and selection: the public goods hypothesis. In *The Pangenome: Diversity, Dynamics and Evolution of Genomes* (Tettelin, H. and Medini, D., eds), pp. 151–167, Springer International Publishing
33. Maddamsetti, R. and Lenski, R.E. (2018) Analysis of bacterial genomes from an evolution experiment with horizontal gene transfer shows that recombination can sometimes overwhelm selection. *PLoS Genet.* 14, e1007199
34. Domingo-Sananes, M.R. and Mclnerney, J. (2019) Selection-based model of prokaryote pangenomes. *bioRxiv* Published online October 21, 2019. https://doi.org/10.1101/782573
35. van Dijk, B. *et al.* (2020) Slightly beneficial genes are retained by bacteria evolving DNA uptake despite selfish elements. *eLife* 9, 1–36
36. Brito, P.H. *et al.* (2018) Genetic competence drives genome diversity in *Bacillus subtilis*. *Genome Biol. Evol.* 10, 108–124
37. González-Torres, P. *et al.* (2019) Impact of homologous recombination on the evolution of prokaryotic core genomes. *mBio* 10, e02494-18
38. Faure, G. *et al.* (2019) CRISPR–Cas: complex functional networks and multiple roles beyond adaptive immunity. *J. Mol. Biol.* 431, 3–20
39. Sela, I. *et al.* (2016) Theory of prokaryotic genome evolution. *Proc. Natl. Acad. Sci. U. S. A.* 113, 11399–11407
40. Niehus, R. *et al.* (2015) Migration and horizontal gene transfer divide microbial genomes into multiple niches. *Nat. Commun.* 6, 1–9
41. Baumdicker, F. and Pfaffelhuber, P. (2014) The infinitely many genes model with horizontal gene transfer. *Electron. J. Probab.* 19, 1–28
42. Beceiro, A. *et al.* (2013) Antimicrobial resistance and virulence: a successful or deleterious association in the bacterial world? *Clin. Microbiol. Rev.* 26, 185–230
43. San Millan, A.S. *et al.* (2014) Positive selection and compensatory adaptation interact to stabilize non-transmissible plasmids. *Nat. Commun.* 5, 5208
44. Milner, D.S. *et al.* (2019) Environment-dependent fitness gains can be driven by horizontal gene transfer of transporter-encoding genes. *Proc. Natl. Acad. Sci. U. S. A.* 116, 5613–5622
45. Levin, B.R. *et al.* (1988) Frequency-dependent selection in bacterial populations [and discussion]. *Philos. Trans. R. Soc. Lond. Ser. B Biol. Sci.* 319, 459–472
46. Corander, J. *et al.* (2017) Frequency-dependent selection in vaccine-associated pneumococcal population dynamics. *Nat. Ecol. Evol.* 1, 1950–1960
47. McNally, A. *et al.* (2019) Diversification of colonization factors in a multidrug-resistant *Escherichia coli* lineage evolving under negative frequency-dependent selection. *mBio* 10, e00644-19
48. Morris, J.J. *et al.* (2012) The Black Queen Hypothesis: evolution of dependencies through adaptive gene loss. *mBio* 3, e00036-12
49. Fullmer, M.S. *et al.* (2015) The pan-genome as a shared genomic resource: mutual cheating, cooperation and the black queen hypothesis. *Front. Microbiol.* 6, 1–5
50. Koskiniemi, S. *et al.* (2012) Selection-driven gene loss in bacteria. *PLoS Genet.* 8, 1–7
51. Mira, A. *et al.* (2001) Deletional bias and the evolution of bacterial genomes. *Trends Genet.* 17, 589–596
52. Sela, I. *et al.* (2018) Estimation of universal and taxon-specific parameters of prokaryotic genome evolution. *PLoS One* 13, e0195571
53. Seth, E.C. and Taga, M.E. (2014) Nutrient cross-feeding in the microbial world. *Front. Microbiol.* 5, 350
54. Whelan, F.J. *et al.* (2020) Coinfinder: detecting significant associations and dissociations in pangenomes. *Microb. Genomics* 6, e000338
55. Whelan, F.J. *et al.* (2020) Evidence for selection in a prokaryote pangenome. *bioRxiv* Published online October 28, 2020. https://doi.org/10.1101/2020.10.28.359307
56. Cui, Y. *et al.* (2020) The landscape of coadaptation in *Vibrio parahaemolyticus*. *eLife* 9, 1–23
57. Press, M.O. *et al.* (2016) Evolutionary assembly patterns of prokaryotic genomes. *Genome Res.* 26, 826–833
58. Cohen, O. *et al.* (2012) Uncovering the co-evolutionary network among prokaryotic genes. *Bioinformatics* 28, i389–i394
59. Mell, J.C. and Redfield, R.J. (2014) Natural competence and the evolution of DNA uptake specificity. *J. Bacteriol.* 196, 1471–1483
60. Evans, B.A. and Rozen, D.E. (2013) Significant variation in transformation frequency in *Streptococcus pneumoniae*. *ISME J.* 7, 791–799
61. Bolotin, E. and Hershberg, R. (2016) Bacterial intra-species gene loss occurs in a largely clocklike manner mostly within a pool of less conserved and constrained genes. *Sci. Rep.* 6, 35168
62. Sela, I. *et al.* (2019) Selection and genome plasticity as the key factors in the evolution of bacteria. *Phys. Rev. X* 9, 031018
63. Kuo, C.-H. and Ochman, H. (2009) Deletional bias across the three domains of life. *Genome Biol. Evol.* 1, 145–152
64. Raynes, Y. and Sniegowski, P.D. (2014) Experimental evolution and the dynamics of genomic mutation rate modifiers. *Heredity (Edinb)* 113, 375–380
65. Couce, A. *et al.* (2017) Mutator genomes decay, despite sustained fitness gains, in a long-term experiment with bacteria. *Proc. Natl. Acad. Sci. U. S. A.* 114, 1705887114
66. Iranzo, J. *et al.* (2017) Disentangling the effects of selection and loss bias on gene dynamics. *Proc. Natl. Acad. Sci. U. S. A.* 114, E5616–E5624
67. Sung, W. *et al.* (2012) Drift-barrier hypothesis and mutation-rate evolution. *Proc. Natl. Acad. Sci. U. S. A.* 109, 18488–18492
68. Ramiro, R.S. *et al.* (2020) Low mutational load and high mutation rate variation in gut commensal bacteria. *PLoS Biol.* 18, e3000617
69. Kimura, M. and Crow, J.F. (1964) The number of alleles that can be maintained in a finite population. *Genetics* 49, 725–738
70. Hamilton, M.B. (2009) *Population Genetics*, Wiley-Blackwell, p. 407
71. Tenaillon, O. *et al.* (2001) Second-order selection in bacterial evolution: selection acting on mutation and recombination rates in the course of adaptation. *Res. Microbiol.* 152, 11–16
72. West, S.A. *et al.* (2006) Social evolution theory for microorganisms. *Nat. Rev. Microbiol.* 4, 597–607