Trends in Microbiology

Mechanisms that shape microbial pangenomes

--Manuscript Draft--

Manuscript Number:	TIMI-D-20-00249R3
Article Type:	Opinion
Keywords:	pangenomes; genome evolution; natural selection; genetic variation
Corresponding Author:	Maria Rosa Domingo-Sananes University of Nottingham Nottingham, UNITED KINGDOM
First Author:	Maria Rosa Domingo-Sananes
Order of Authors:	Maria Rosa Domingo-Sananes
	James O. McInerney
Abstract:	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.

University of Nottingham
UK | CHINA | MALAYSIA

School of Life Sciences
University of Nottingham
University Park
Nottingham
NG7 2RD

09 December 2020

Dr. Shankar Iyer
Editor
Trends in Microbiology

Dear Shankar Iyer,

Thank you for the opportunity to submit a revised version of our manuscript, titled "Mechanisms that shape microbial pangenomes" to Trends in Microbiology.

Thank you again for your comments, which have been very helpful for improving and polishing our manuscript. We have clarified the requested sentences and addressed all the comments provided in your decision letter. The changes made are highlighted with comments on the revised manuscript file with tracked changes. We also highlight them after this letter. As requested, we are also submitting a clean version of the manuscript without tracked changes and comments.

Thank you very much for your consideration.

Sincerely,

Maria Rosa Domingo-Sananes and James McInerney

For correspondence: Maria.DomingoSananes@nottingham.ac.uk,
James.McInerney@nottingham.ac.uk

***Sentence(s) in manuscript**: “Another key example has been inferred in Campylobacter jejuni, where the presence of a seven-gene region is associated with host preference and not phylogeny”*
***Editor comment/ suggestion**: An example in itself cannot be inferred but rather a specific aspect is based on observations. Please revise this sentence accordingly. I would suggest: “Another key example involves the bacterium, Campylobacter jejuni, where the presence of a seven-gene region is associated with host preference and not phylogeny”*

> Changed as suggested to: “Another key example involves the bacterium Campylobacter jejuni"

***Sentence(s) in manuscript**: “Furthermore, these relationships could also occur between accessory genes and particular sequence variants in core genes or accessory genes, or between accessory genes and general genetic background”*
It is unclear what elements of the general genetic background accessory genes are being compared to, so please revise this.

> Changed to: "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.”

<u>Comment 2</u>*: The phrase “core genes or accessory genes” begs the question of what the subject specifically is here: interactions of sequence variants between core and accessory genes (OR) interactions of sequence variants between different core genes and different accessory genes. I would suggest that the authors elaborate this entire sentence so the relationships that are intended to be conveyed better stand out, even if this means addition of another sentence.*

> Change 2: "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.”

***Sentence(s) in manuscript**: “Additionally, gene loss is known to vary between species [62–64]. Therefore, there is variation in overall rates of gene gain and loss between and within species. Theoretically, if there is variation in the rates of gene gain and loss between individuals, it could be acted upon by selection [8]”*
***Editor comment/ suggestion**: The occurrence of “therefore” in the second sentence above is not entirely justified, given the previous few sentences. I would suggest the authors to rather summarize the contents of the section by inferring that variation exists both between and within species (stating: individuals within the same species), before concluding with the last sentence. These bridges will aid readability and help readers follow the flow of the manuscript.*

> Changed to: "Additionally, gene loss is known to vary between species [62–64]. 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].”

***Sentence(s) in manuscript**: “On the other hand, in an attempt to draw a parallel with NSV…”*

***Editor comment/ suggestion**: It is unclear what exactly is being referred to here, while attempting to draw a parallel to NSV. I would suggest that the authors explicitly state this, even if this would involve reiteration.*

> Changed to: "On the other hand, 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.

***Sentence(s) in manuscript**: "Interestingly, adding a third class of genes, such that the pangenome, which contains genes with high, intermediate and no mobility, improves the fit even further [24]. These gene classes might correspond respectively to the proposed classes of cloud, shell and core genes, which may have decreasing levels of mobility [7,24]."*
***Editor comment/ suggestion**: The very first sentence, in addition to being incomplete, introduces a "third class", while the previous sentences do not explicitly demarcate the other classes. Furthermore, the correspondence of "cloud, shell and core genes" is not with respect to the gene classes but rather with the levels of mobility outlined in the previous sentence. These sentences, taken together are confusing as to correspondence and the classes of genes, itself so I would suggest the authors to carefully revise this or even possibly combine these (to more explicitly establish the correspondence) such that the intended meaning is unchanged.*

> Thank you for clarifying this, and apologies for not addressing your comment properly before. I think I now understand what you meant. The last part of this paragraph has be changed to:
>
> "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), since this simulates the presence of the fraction of essential, common genes that the model with a single gene class has trouble predicting [24]. This reflects the importance of selection in maintaining core genes, which is perhaps unsurprising. This model with two gene classes, highly mobile and immobile (essential) genes, fits some real data vary well, such as the gene accumulation curves for Streptococcus pneumoniae [24]. Interestingly, adding a third class of genes with intermediate rates of gene gain and loss (and thus motility), improves the fit even further [24]. These gene classes of high, medium and no motility 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]. This may also imply that the "shell" genes, with intermediate mobility, may correspond to genes that are maintained to an extent by selection and through some of the mechanisms described in the main text."

<u>Comment 2 (form word file):</u> *Multiple comments on this section: simplification, clarification, use of motility and further explanation of fit to the model*

> Change 2: Further improvement can be achieved by explicitly incorporating core genes (genes that cannot be gained or lost). This is because incorporating a second class of immobile genes simulates the fraction of essential genes in the genome [24]. This reflects the importance of

selection in maintaining core 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 Streptococcus pneumoniae, that is the number of addition genes found with each additional genome that is analysed [24]. However, the fit is less accurate for the core-gene depletion curve, where the model over-estimates 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 trough selection and the other mechanisms described in the main text.

***Sentence(s) in manuscript**: "Describes the how the fitness effect of a genetic variant (gene) can depend on the environment"*
Editor comment/ suggestion: change to "Describes how the fitness..."

Thank you for spotting this. Edited as suggested

Mechanisms that shape microbial pangenomes

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

September 2020

[1]School of Life Sciences, University of Nottingham, UK

*Correspondence: Maria.DomingoSananes@nottingham.ac.uk (MR Domingo Sananes); James.McInerney@nottingham.ac.uk (JO McInerney)

Keywords

Pangenomes, genome evolution, natural selection, genetic variation

Abstract

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.

Abbreviations

NSV, nucleotide sequence variation; GCV, gene content variation; DFE, distribution of fitness effects; NFDS, negative frequency-dependent selection.

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 single nucleotide polymorphisms (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 non-coding 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** (Figure 1e, Glossary) have been proposed to better quantify pangenome diversity [11].

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 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] (Box1). 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 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).

On the other hand, 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, 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. On the other hand, 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 vary 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.

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 gain/loss balance may be much more important in GCV compared to 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 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 in 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 benefits to cells growing in high glycerol concentrations, but was deleterious for cells growing in glucose [44]. Another key example involves the bacterium *Campylobacter jejuni*, where 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 carrying 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 changes. 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 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 it 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 the 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 interactions may be rare, they have been detected, for example as metabolic cross-feeding, where individuals exchange metabolites that benefit both 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 to when each gene is present on its own (that is, non-additive contributions to fitness). The fitness effect of being jointly present might be an overall positive or a negative effect. These types of interactions 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 is 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 interactions may occur together. Further analysis should focus on dissecting how common all these types of interactions 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 distributions 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 *Streptococcus 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 [37,61]. Additionally, gene loss is known to vary between species [62–64]. 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]. On the other hand, 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[65]. However, elevated mutation rates can also result in the accumulation of many other neutral and slightly deleterious mutations [66]. 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)

[66]. 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, [52,62,64]. 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) [39,51,64,67]. The extent of this loss bias varies between organisms [52,64], 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 [68]. 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 [69], recombination in prokaryotes may be sufficiently low for second-order selection [65,66]. 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, *i.e.* 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. 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.

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, where 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 to 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 [70]. As in the previous case, the original formulation of the model assumes that every gene can only be acquired 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 *Streptococcus 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 over-estimates 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 trough selection and the other mechanisms described in the main text.

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]):

$$\frac{dx}{dt} = r_g(1-x) + sx(1-x) - r_l x$$ Equation 1

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 (Figure 2a-d, [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 take into account 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 (Box 2). 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 for described by the model. Different lines show different combinations of rates of gene gain and loss.

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 maintain 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 [71].

- **Negative frequency-dependent selection**: 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 becomes more common in the population. This mechanism can cause fast divergence between populations of the same species, and thus contribute to increased variation.

- **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, where 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**: 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].

- **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 [72].

- **Social interactions**: Describes interactions between organisms of the same or different species that can affect their fitness. Interactions can be mutually beneficial, altruistic, selfish or spiteful [73].

Acknowledgements

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.* 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, 1574–1574

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 (80-.).* 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 (80-.).* 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.* 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 McInerney, 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 McInerney, J. (2019) Selection-based model of prokaryote pangenomes. *bioRxiv* DOI: 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.* 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.* 116, 5613–5622

45 Levin, B.R. *et al.* (1988) Frequency-Dependent Selection in Bacterial Populations [and Discussion]. *Philos. Trans. R. Soc. Lond. 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* DOI: 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, 826–33

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 Faure, G. *et al.* (2019) CRISPR–Cas: Complex Functional Networks and Multiple Roles beyond Adaptive Immunity. *J. Mol. Biol.* 431, 3–20

62 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

63 Sela, I. *et al.* (2019) Selection and Genome Plasticity as the Key Factors in the Evolution of Bacteria. *Phys. Rev. X* 9,

64 Kuo, C.-H. and Ochman, H. (2009) Deletional Bias across the Three Domains of Life. *Genome Biol. Evol.* 1, 145–152

65 Raynes, Y. and Sniegowski, P.D. (2014) Experimental evolution and the dynamics of genomic mutation rate modifiers. *Heredity (Edinb).* 113, 375–380

66 Couce, A. *et al.* (2017) Mutator genomes decay, despite sustained fitness gains, in a long-term experiment with bacteria. *Proc. Natl. Acad. Sci.* DOI: 10.1073/pnas.1705887114

67 Iranzo, J. *et al.* (2017) Disentangling the effects of selection and loss bias on gene dynamics. *Proc. Natl. Acad. Sci.* 114, E5616–E5624

68 Sung, W. *et al.* (2012) Drift-barrier hypothesis and mutation-rate evolution. *Proc. Natl. Acad. Sci.* 109, 18488–18492

69 Ramiro, R.S. *et al.* (2020) Low mutational load and high mutation rate variation in gut commensal bacteria. *PLOS Biol.* 18, e3000617

70 Kimura, M. and Crow, J.F. (1964) The number of alleles that can be maintained in a finite population. *Genetics* 49, 725–38

71 Hamilton, M.B. Population Genetics. . (2009) , Wiley-Blackwell, 407

72 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

73 West, S.A. *et al.* (2006) Social evolution theory for microorganisms. *Nat. Rev. Microbiol.* 4, 597–607

Figure Legends

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 (solid lines) and core-gene depletion curves (dotted 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.

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 Box1). (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 Box1). (e) Schematic examples of interactions that contribute to gene content variation: Gray 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 of the presence of gene *M* (yellow); the presence of gene *P* (purple) is conditional on the absence of gene *M* (yellow).

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.

Formatted: Indent: Left: 0.38", No bullets or numbering

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 lifestyle, environment and/or taxonomic properties?
- What determines loss rates and genome sizes? Does selection play a role?

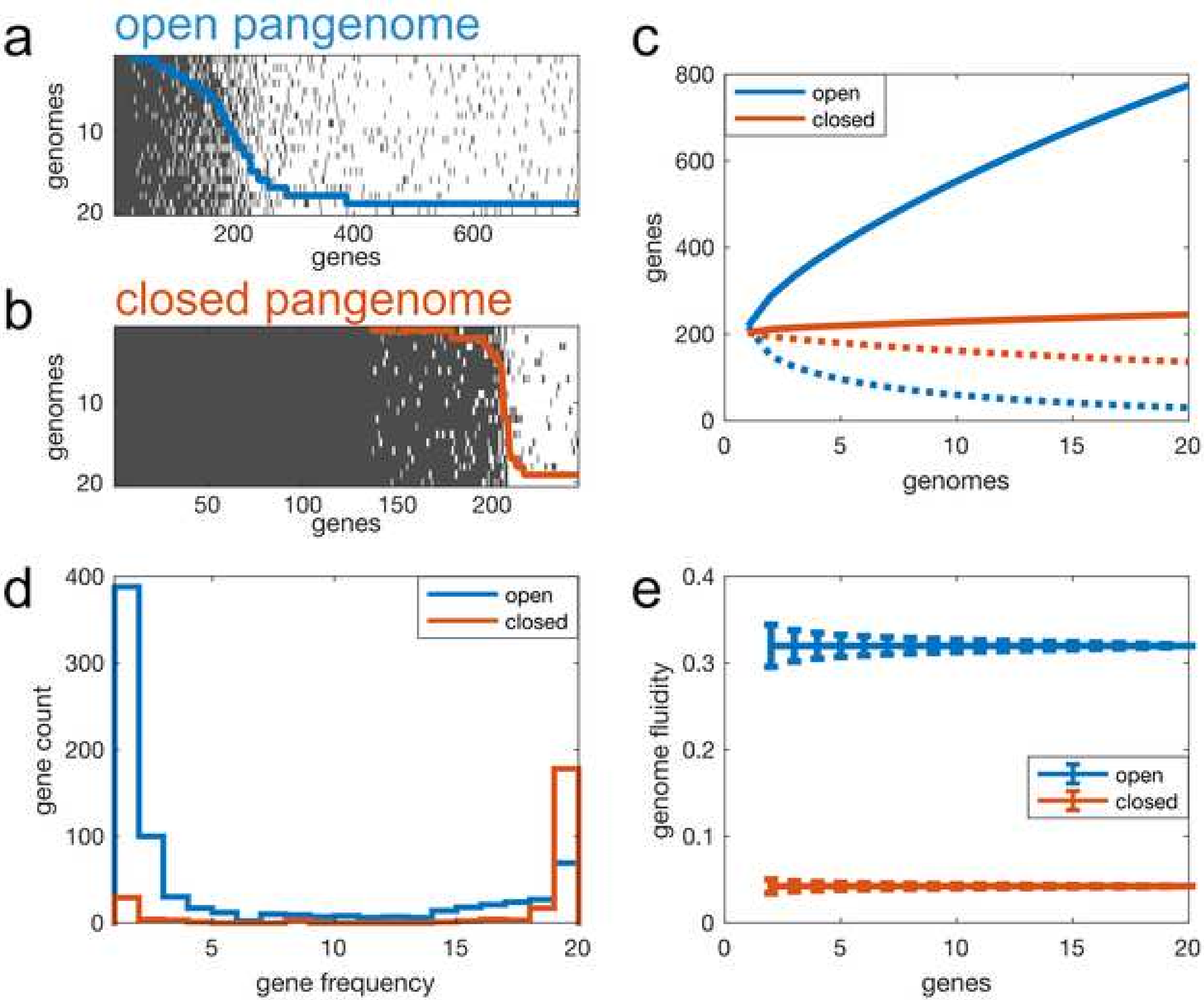

a
open pangenome
genomes
10
20
200
400
600
genes
b
closed pangenome
genomes
10
20
50
100
150
200
genes
c
800
600
400
200
0
genes
open
closed
0
5
10
15
20
genomes
d
400
300
200
100
0
gene count
open
closed
5
10
15
20
gene frequency
e
0.4
0.3
0.2
0.1
0
genome fluidity
open
closed
0
5
10
15
20
genes

Figure 2

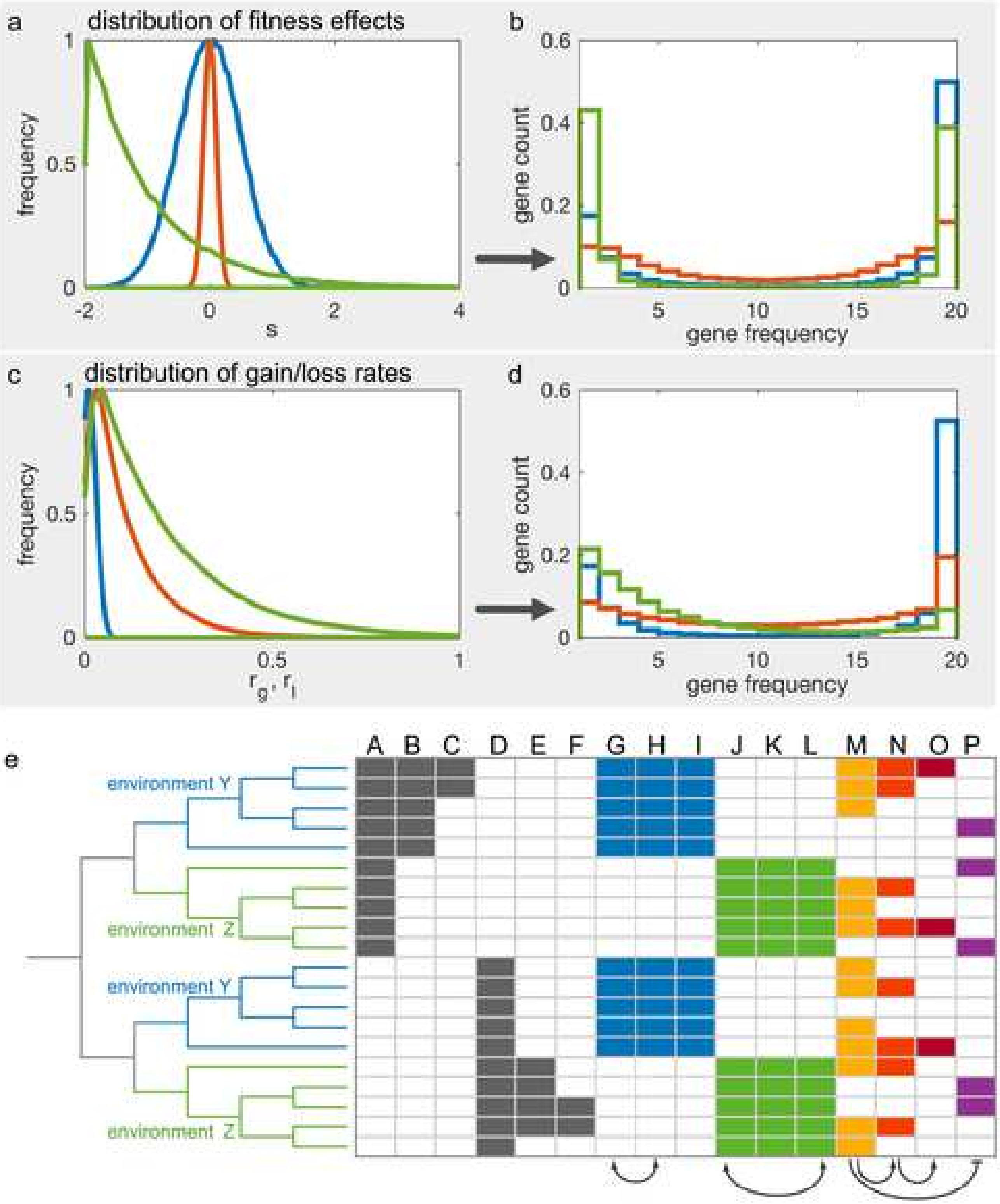
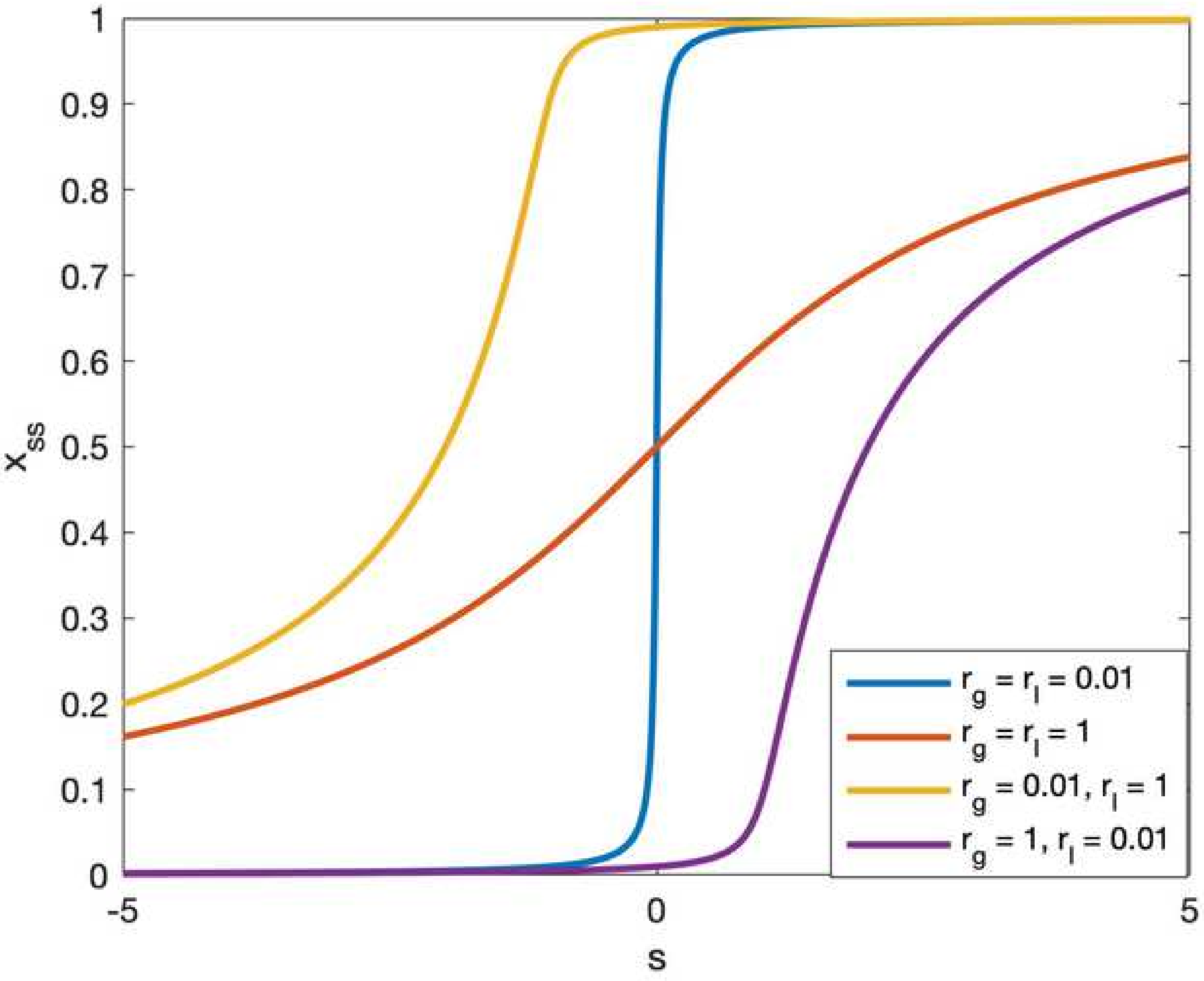

a
distribution of fitness effects
frequency
1
0.5
0
-2
0
2
4
s
b
gene count
0.6
0.4
0.2
0
5
10
15
20
gene frequency
c
distribution of gain/loss rates
frequency
1
0.5
0
0
0.5
1
r_g, r_l
d
gene count
0.6
0.4
0.2
0
5
10
15
20
gene frequency
e
A B C D E F G H I J K L M N O P
environment Y
environment Z
environment Y
environment Z

Figure I


1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
x_{ss}
-5
0
5
s
$r_g = r_l = 0.01$
$r_g = r_l = 1$
$r_g = 0.01$, $r_l = 1$
$r_g = 1$, $r_l = 0.01$

Response to the Editor and Reviewers

Our reply to each comment and suggestion is below. The original comments are in in blue italics while our reply is in black.

Comments from the Editor:

"One of the biggest concerns remains the substantial overlap between the discussions in the current manuscript and other published work. For instance, the discussions regarding the importance of selection (as Reviewer #2 points out). Given this concern and also additional overlap with the 2018 article by Eduardo P C Rocha, I would recommend that the authors revise the manuscript to better flesh out its novelty and if possible, also provide a comprehensive list of key aspects that are novel to this manuscript and what percentage of the manuscript comprises of said novel aspects"

Thank you for taking the time to read and our manuscript. We appreciate your insightful comments and suggestions. While we agree that some issues presented in our manuscript have been discussed before, we think that these issues are not resolved, which is one of the main points we make. We believe part of the issues concerning overlap, are limited to the first, introductory section. We have made this section more concise, but we still have to mention these points to introduce the reader and present these evolutionary and population genetics ideas to the wide audience of Trends in Microbiology. We have also extensively edited the manuscript to clarify and flesh out the important and novel points in our manuscript and have also included them more explicitly in the Highlights section.

In terms of the overlap with discussions on the importance of selection, we believe that this is still an unresolved issue. We mention this discussion in our introduction section, which is important for the context of our article. We argue that because current neutral models do not quite fit the real-world data, we need to consider mechanisms that promote or maintain gene content diversity. Proposing a set of such mechanisms is the key point of our article, which we develop in the three sections after the introduction.

We believe our article does not overlap but partially builds on some of the ideas presented by Rocha in 2018. The similarities we perceive and how we have addressed them are as follows:

- Both manuscripts highlight some of the difficulties in trying to understand genetic variation and adaptation in prokaryotes. We think this is a point that is worth repeating in the context of our article and has been also mentioned by other authors (e.g., see manuscript references 14 and 15). However, our perspective differs in our focus on gene content variation, whereas Rocha 2018 focuses on sequence variation more strongly. Crucially, we highlight some important differences between GCV and NSV, which is a key point from our perspective. In particular, we propose that variation in rates of gene gain and loss for different genes and species may have a large impact on pangenome diversity, and potentially lead to some genes being maintained at intermediate frequencies due to gain/loss balance.

- Both articles mention the need for theoretical approaches to account for particular features of prokaryote populations. While Rocha 2018 focuses on the need to account for strong linkage and the infinitely many genes model, we believe that other issues need to be considered too. In particular the variation in gene gain/loss rates and fitness effects, along with the mechanisms and interactions we propose.

- Both articles mention the need of approaches to identify the proportion and identities of beneficial accessory genes. We believe that this is a general conclusion and aim of the analyses of gene content diversity.

- Both articles mention the possibility of second order selection on transformation rates. However, we go further and argue that second order selection could happen on many more mechanisms that determine gene and loss.

There are also many important points raised by Rocha which we do not consider in our article.

We believe our overall perspective is novel, and that at least 50% of our article contains novel ideas. Our main focus is to propose, list and explain the many diverse mechanisms that contribute to gene content variation. As far as we are aware, this has not been scrutinised before.

As mentioned previously, the first section (Pangenomes and why they matter) is mostly introductory, and therefore it does not have novel ideals. The novelty in section two (Parameters that shape pangenomes: fitness effects and rates of gene gain and loss) comes from explicitly considering potential differences in the distributions of fitness effects between species and cells, and even more importantly proposing the existence of extensive variation in the rates of gene gain and loss (our proposal of the existence of distributions for these parameters) and variation between species. We also propose that gain/loss balance is a mechanism that could maintain genes at stable intermediate frequencies if the rates of gain and loss for particular genes are high.

In section three (Interactions that shape pangenomes), we explicitly propose gene-by-environment and gene-by-gene interactions as mechanisms that could maintain GCV. Negative frequency-dependent selection [45] and Black Queen dynamics [48, 49] have been proposed previously. However, we propose the possibility that positive frequency-dependent selection and Black-Queen-like interactions unrelated to genome reduction could also have a role.

For section four (The evolvability of pangenomes), while as mentioned above the idea of second-order selection on transformation rates has been previously mentioned (Rocha 2018), we propose that this could apply in general for mechanisms that influence gene gain (including mechanisms that *prevent* gene gain), and for mechanisms that could influence gene loss.

"While the overall premise and the expertise of the authors is very clearly compelling, I want to make sure that majority of the discussions also offer unique perspectives. On the subject of perspectives, I have a slight qualm regarding how the "Opinion"-nature of the current piece

as it does not offer a clearly unique opinion or hypothesis. I would love to hear your thoughts on all of these comments."

We believe the overall picture we propose is novel, in particular our focus on the diversity of mechanisms that can contribute to gene content variation. However, it is true that some of these ideas have been separately presented previously. Although we would much favour this article remaining an opinion piece, we would be open to the idea of publishing as a review article instead.

In addition to the comments made by the reviewers, I have also included my own minor comments/ suggestions in an annotated version of the manuscript. Overall, would suggest that the authors include a few more examples (of genes/ systems across prokaryotes), wherever possible and also, be more mindful of the scope of the journal and clarify the specificity to prokaryotes in certain instances.

Thank you again for your comments in the manuscript. We have itemised them here as well:

1. Given the mention of GCV, "especially in prokaryotes" and the occurrence of the highlighted phrase, given the scope of the journal, I would suggest that the authors demarcate and mention that the discussions/perspectives in this piece would revolve around prokaryotes. It might also be worth mentioning possible overlap between the two, if the authors feel it necessary.

Thank you for this suggestion. We made this point prominent at the end of paragraph 1 in section 1. The last few sentences in that paragraph now read:

"Pangenomes contain core genes present in all individuals and accessory genes whose presence varies. The concept has been expanded to consider structural and copy number variation in both protein-coding and non-coding 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]. Some of the mechanisms described here can apply to eukaryotes and higher taxonomic levels, in this article we focus mostly on GCV in prokaryote species"

We have also made the title reflect our focus on microbial pangenomes.

2. This is a great example but I wonder if this part would benefit from a few other examples of gene (families)?

This is a good suggestion. We have added an explicit example of gene-by-environment interactions for antibiotic resistance in *Pseudomonas aeruginosa*, and another example for metabolite transporter genes in fungi and *Saccharomyces cerevisiae* (page 6).

3. The number of examples where specific species are instanced are quite low throughout the manuscript. These usually help better place the summary/ perspective within the context of the field. I would suggest that the authors include a few more specific examples of specific genes/ interactions, where possible, across other species.

Again, thank you for this suggestion. We had not incorporated many exampled previously partly due to space concerns. We agree that this helps to place the article in the context of the filed. We have incorporated examples throughout the manuscript:

- Example of measured differences in the fitness effects of accessory genes in E. coli (page 3)
- Prochlorococcus niche-associated genes as an example of how GCV may contribute to adaptation (page 3)
- Transposon-associated colistin resistance as an example of a gene with high rate of gain (page 5)
- Longitudinal analyses of pangenomes in *E. coli* and *Streptococcus pneumoniae* which suggest that negative frequency-dependent selection is common in these organisms (page 7).
- Mention marine oligotrophic bacteria with small genomes (*Prochlorococcus* and "*Candidatus* Pelagibacter,"), which may be involved in Black Queen dynamics
- Associations between niches, SNPs and accessory genes in *Vibrio parahaemolyticus*.
- Mention variability in transformation efficiencies between strains in *S. pneumoniae.*

Reviewer's Comments

Reviewer #1:

I found this review on bacterial pan genomes informative and to-the-point. For someone new to the field, it covers all the basics (nicely illustrated as well in Figure 1), but it also gives insights into the wider discussion on what evolutionary forces underlie variation in pan genome sizes. This latter part discusses what we know of the rate and fate of gene gains and losses, how interactions between the genome and the environment and between genes in genomes could result in selection for accessory genome diversity and how there could be selection for not just the presence of accessory genes, but selection for mechanisms that regulate acquiring accessory genes.

Thank you for taking the time to read and comment on our manuscript. We very much appreciate the nice comments and summary. We have aimed to provide an introduction to the field and evolutionary aspects of pangenomes. We do think that our manuscript goes beyond a review in that it proposes a general perspective into how pangenome diversity might be maintained.

I have no major comments, but have some niggles below both on language and on a few parts which could be explained better and could be expanded:

1. Could the authors rethink the title? What is the difference between a 'parameter' and a 'process' in the context of shaping pan genomes? It is simultaneously technical and vague sounding.

We take on board this suggestion and the potential issues with the original title. We have changed it to a more generic: "Mechanisms that shape microbial pangenomes"

2. "important for predicting and understanding the evolution of virulence, pathogenicity, and the spread of antimicrobial resistance." Virulence and pathogenicity seem to cover the same thing. Are there other fields where pan genome evolution could matter, bioremediation or industrial applications perhaps? (I noticed biotechnology is mentioned at the end of the paper).

Thank you for these suggestions. We refer to pathogenicity as the ability of an organism to infect a host, while virulence reflects the severity of disease caused by a pathogen. Therefore, they refer to different properties of a pathogen that may be influenced by different genes. However, to make the point clearer, we have replaced pathogenicity with infectiveness. We have also added a couple of sentences to highlight the importance of GCV for biotechnology. This part of the manuscript now reads:

"Understanding how GCV contributes to adaptation is not only interesting from an evolutionary perspective, but it is important for predicting and understanding the evolution of virulence, infectiousness, and the spread of antimicrobial resistance in pathogens. 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 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"

3. Use of "at vastly different rates" repetitive.

Thank you for pointing this out. The repetition has been changed to:

"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 at different frequencies."

4. It would be good to explain the drift barrier hypothesis.

This is a very good suggestion. We have added an explanation to the drift barrier hypothesis, which now reads:

"Another possibility is the existence of a "drift barrier" similar to that proposed for the evolution of mutation rates [69]. 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"

5. BQH definition in Glossary: I suggest "due to the need to complete functional processes, which become partially encoded in different cells" to be rephrased.

Thank you for this suggestion. We have changed the definition in the glossary to make this term clearer:

"**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 maintain genes at stable intermediate frequencies."

6. I think the BQH explanation in the main text also can be improved. It is not very clear from the text that although cross-feeding has been demonstrated that this has not been linked to genome size in these studies. The phrasing (collaborate, complex tasks) is a bit off as well.

Thank you for pointing this out. We did not mean to imply a direct relationship between cross-feeding and genome size. We just wanted to point out that the types of interactions proposed to drive Black Queen dynamics do happen in nature in the form of cross-feeding interactions. We hope we have now made this clearer by changing the text to:

"A further mechanism that has been proposed to contribute to GCV is the distributed genome or Black Queen hypothesis [48,49]. The idea is that "leaky" functions, such as 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 that function. This may have the benefit of allowing organisms to maintain a smaller genome [50], as has proposed for some oligotrophic marine bacteria, such as Prochlorococcus and "Candidatus Pelagibacter," [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 the smaller genome may not be the direct benefit [19,39,51,52], additional factors could encourage interactions similar to those proposed by the Black Queen hypothesis, such as completing tasks in the most appropriate genetic background or division of labour. While these types of complex social interactions may be rare, they have been detected, for example in the form of cross-feeding, where individuals exchange metabolites that benefit both 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]."

7. The conclusion (rightly) mentions that there are several explanations for pan genomes which are not necessarily mutually exclusive. However, I would like to see these options a bit more formalised (in a table?) or at least discussed in a bit more detail. For instance: "large Ne combined with phylogeny and population structure and constant gain and loss of neutral genes" is quite a lot to take in, and why/how population structure is important is not explained.

Thank you for this suggestion. We have added a small figure to highlight the different mechanisms and parameters that can contribute to pangenome diversity (Figure 3). We have also edited this part of the conclusion to make it more readable. It now reads:

"There are several different explanations for GCV and intermediate gene frequencies. For example, largely 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. We summarise these factors and their likely relationship with pangenome diversity (genome fluidity) in Figure 3"

8. "To accomplish these goals, we need to develop a testable theoretical framework that can capture the processes and mechanisms that we have considered, followed by testing these models with empirical data and experiments." I agree, but can the authors be a little bit more specific about possible ways forward?

This is a very good suggestion. We have added a few sentences to propose possible future research pathways:

"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 effect of variation in gene fitness and gain and 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 yet more whole genomes with associated metadata. Direct observation of pangenome evolution from longitudinal studies and experimental evolution will also help us to disentangle the mechanisms that shape microbial pangenomes."

9. Box 1 "it does not entirely capture these features when fit to real data." That is right, but this could be explained, e.g. genome size DOES vary between strains in a species, and gene transfers can be from within the population (in fact, most likely are, as recombination with homologous flanking DNA will greatly increase recombination efficiency compared to the original introduction from a divergent donor).

Thank you for this comment. We have added a few sentences to explain how real data does not quite fit the Haegeman & Weitz model, and how the fit can (and may be improved):

"However, it does not entirely capture these features when fit to real data. 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 a species, and gene transfer can occur within a population, incorporating these features might improve the explanatory power of this model."

10. "Further improvement can be achieved by explicitly incorporating core genes" why?

We have added an explanation for how including a core gene category improves the fit of the IMG model to real data:

"Further improvement can be achieved by explicitly incorporating core genes (genes that cannot be gained or lost), since this simulates the presence of the fraction essential, common genes that the model with a single gene class has trouble predicting."

11. I think it would be good to explain the axes for Figure I in its Legend.

Thank you for this suggestion. We have added an extended description to the legend for this figure. It now reads:

"Figure I. Gene frequency in a population with respect to fitness effect for different rates of gene gain and loss, according to the model described in Box 2. The x-axis represents the contribution of a gene to fitness to the cell, while the y-axis indicates the expected frequency of genes with the corresponding fitness value. This is the frequency that would be expected exclusively under gain-loss balance."

Reviewer #2:

The authors have submitted an opinion manuscript about the parameters and processes that shape pangenomes.

In their manuscript, they highlight various possible evolutionary mechanisms that can influence the distribution of genes in pangenomes and argue that we need novel theoretical approaches and empirical data to identify the important mechanisms among them.

The points made are similar to the argumentation in the article:

Rocha, E. P. C. (2018). Neutral Theory, Microbial Practice: Challenges in Bacterial Population Genetics. Molecular Biology and Evolution, 35(6), 1338-1347. https://doi.org/10.1093/molbev/msy078

with a stronger focus on pangenomes and selection.

Overall the manuscript highlights different mechanisms, but it is made clear that the authors favor selection as the most important one. In addition, the authors have added two paragraphs on interactions that shape pangenomes and the evolvability of pangenomes which raise important points.

These two paragraphs are the strongest part of the manuscript and I enjoyed reading these parts. The first two paragraphs are less inspiring as arguments for the importance of selection of previous publications by the authors and others are repeated and the reasoning with regard to the comparison to existing neutral models is partly vague and imprecise.

However, as this is an opinion article this might be perfectly fine?

Thank you for taking the time to read and comment on our manuscript. We agree that arguing for the need of novel theoretical approaches and empirical data to understand genetic variability are points in common between our manuscript and the article by Rocha 2018. We also agree that most of the content presented in the first section has been discussed previously, but we believe that it is necessary as an introduction to the field and issues for a wide audience.

However, we do not entirely agree that the content of the second section has been published previously. To address this issue, we have tried to clarify some of the novel points in this section, including the consideration of diversity in the rates of gene gain and loss, and the possibility of gene gain/loss balance as a mechanism that contributes to gene content diversity. We have also aimed to make our discussion of neutral models more precise, while maintaining readability for a wide audience.

We are very grateful for the nice comments on the final two sections of the manuscript.

1. page 2: the development of population genetics was not a "resulting" from nucleotide sequence variation studies. Population genetics is much older but has of course been used to study NSV in the last decades. I would rephrase the first sentence.

Thank you for this observation which we agree with. We have changed this sentence to:

“Nucleotide sequence variation (NSV) in populations, in the form of single nucleotide polymorphisms (SNPs) and short indels, has been the subject of study for decades, and analysed trough the robust theoretical framework of population genetics, which aims to characterise and model genetic variation”

2. page 2: I did not understand why "we still lack a robust theoretical framework". There are theoretical models as shown in the manuscript. What do the authors consider to be a "robust" framework?

Thank you for this comment. To clarify this issue, we have removed this sentence form the first section. We turn to this point more specifically in the two boxes describing theoretical approaches. We have added a more detailed description of what future models should aim to include to develop a more complete description of GCV and its evolution, in our opinion.

3. page 3: We show here how different mechanisms .... --> better: We highlight here how different mechanisms ... (as this is not a research article)

This is a fair criticism, thank you. We have changed this sentence to: “We describe here how different mechanisms…”

4. page 4: "indicating that selection has a role in shaping pangenomes". if the data does not fit this does not imply selection as the solution. Better: "indicating that selection might have a role in shaping pangenomes"

Again, this is a fair comment our wording. We have made the changed suggested.

5. page 5: The authors should state that the toy model presented in box 2 is only valid for a single gene or genes that are frequently recombining. Otherwise, the genomewide linkage will result in another dynamic. This is exactly the point that has so far hindered building a future model that "bridges the gap".

Thank you for this comment. We have now explicitly mentioned that the model in Box 2 does not take into account linkage. This part now reads:

“Although this simple model can give us some insight, it does not capture the contribution of the evolutionary process or treelike evolution described by the models presented in Box 1. In particular, the model considers genes to be independently gained and lost, and therefore does not take into account genome wide linkage, as do 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."