Tackling the Pangenome Dilemma Requires the Concerted Analysis of Multiple Population Genetic Processes

Franz Baumdicker[1,2,3,*] and Anne Kupczok[4,*]

[1]Cluster of Excellence "Controlling Microbes to Fight Infections", Mathematical and Computational Population Genetics, University of Tübingen, Germany

[2]Cluster of Excellence "Machine Learning: New Perspectives for Science", University of Tübingen, Germany

[3]Institute for Bioinformatics and Medical Informatics (IBMI), University of Tübingen, Germany

[4]Bioinformatics Group, Wageningen University & Research, Netherlands

*Corresponding authors: E-mails: franz.baumdicker@uni-tuebingen.de; anne.kupczok@wur.nl.

Accepted: 19 April 2023

Abstract

The pangenome is the set of all genes present in a prokaryotic population. Most pangenomes contain many accessory genes of low and intermediate frequencies. Different population genetics processes contribute to the shape of these pangenomes, namely selection and fitness-independent processes such as gene transfer, gene loss, and migration. However, their relative importance is unknown and highly debated. Here, we argue that the debate around prokaryotic pangenomes arose due to the imprecise application of population genetics models. Most importantly, two different processes of horizontal gene transfer act on prokaryotic populations, which are frequently confused, despite their fundamentally different behavior. Genes acquired from distantly related organisms (termed here acquiring gene transfer) are most comparable to mutation in nucleotide sequences. In contrast, gene gain within the population (termed here spreading gene transfer) has an effect on gene frequencies that is identical to the effect of positive selection on single genes. We thus show that selection and fitness-independent population genetic processes affecting pangenomes are indistinguishable at the level of single gene dynamics. Nevertheless, population genetics processes are fundamentally different when considering the joint distribution of all accessory genes across individuals of a population. We propose that, to understand to which degree the different processes shaped pangenome diversity, the development of comprehensive models and simulation tools is mandatory. Furthermore, we need to identify summary statistics and measurable features that can distinguish between the processes, where considering the joint distribution of accessory genes across individuals of a population will be particularly relevant.

Key words: pangenome, horizontal gene transfer, selection, neutral evolution.

Significance

Individuals of the same prokaryotic species usually vary in their gene content. This variation results in large pangenomes, that is, the total number of genes in any genome of a species. Different evolutionary processes can create this variation including selective as well as fitness-independent processes. The importance of each of the processes is debated in the literature. Here, we show that it is crucial to distinguish two different modes of horizontal gene transfer and that fitness-independent processes can lead to similar patterns in the data as selective processes. We argue that it is important to consider all the different population genetics processes when modeling and simulating a prokaryotic population.

The Imprecise Application of Population Genetics Models Fuels the Pangenome Debate

Prokaryotic pangenomes show a striking diversity which can even vary within a genus (Brockhurst et al. 2019; Liao et al. 2021). For most prokaryotes, the pangenome contains substantially more genes than any single genome and exceptions are generally intracellular symbionts with limited gene transfer or closely related pathogens (Brockhurst et al. 2019). The evolutionary forces that shape pangenomes—and especially the roles of neutral and adaptive processes—are highly debated (Andreani et al. 2017; McInerney et al. 2017, 2022; Shapiro 2017; Vos and Eyre-Walker 2017). The fundamental pangenome dilemma can be summarized as follows: For prokaryotes, the bioenergetic cost of a whole gene is usually sufficiently large to be perceived by natural selection (Lynch and Marinov 2015) and gene loss is pervasive in bacterial evolution resulting in a deletion bias in genome evolution (Mira et al. 2001). Thus, population genetic theory would predict that genes without adaptive benefits go extinct, whereas beneficial genes become fixed in the population, resulting in pangenomes with only a few accessory genes. However, this is inconsistent with the observation that accessory genes of low and intermediate frequencies are abundant in prokaryotic pangenomes. Notably, due to the effect of random genetic drift, the relative importance of neutral and adaptive processes depends on the effective population size N_e. It has been observed that pangenome diversity increases with larger N_e, resembling the diversity pattern in neutral models (Baumdicker et al. 2012; Andreani et al. 2017).

Nevertheless, this observation could result from various population genetics processes that act on gene frequencies (Box 1, fig. 1). To disentangle these processes, the distinction between the fitness-dependent process selection and all other processes that are fitness-independent is particularly crucial. Importantly, the variants generated by the latter are not necessarily neutral. For example, the emergence of a mutation is fitness-independent, but the mutation might nevertheless result in a variant that can either evolve neutrally or under selection. Despite the multitude of population genetics processes, recent papers only focused on a few of them. For example, based on an eco-evolutionary model for the acquisition of adaptive traits by Niehus et al. (2015), McInerney et al. (2017) suggested that migration between ecological niches is necessary to maintain genome content diversity for selectively advantageous genes. Recently, (McInerney 2022) re-advertised the adaptationist view by emphasizing epistatic interactions between genes. We embrace the recent appraisal that pangenome theory does not currently meet the desiderata for rationality (McInerney 2022), and we hope to reduce the fuzziness of the pangenome concept in this perspective. First of all, the arguments in this discussion strongly rely on insights from Kimura's 50-year-old Neutral Theory, which has been applied abundantly to eukaryotes (Kimura 1983). Notably, when modeling eukaryote population genetics, it is well recognized that genomic variation is jointly shaped by random genetic drift—modulated by N_e and migration—and direct and linked selection; thus, models accounting for multiple factors are essential for accurate inferences (Comeron 2017; Jensen et al. 2019; Johri et al. 2022).

Gene Transfer Processes Differ Between Closely and Distantly Related Bacteria

Eukaryotic species are defined as recombining populations that are reproductively isolated from other species, known as the biological species concept. Thereby, meiotic recombination between homologous chromosomes is included in classical population genetics models (Hartl 2020). In contrast, bacteria are considered haploid and their recombination is not coupled to cell division. Homologous recombination (HR) results in the transfer of genetic material between closely related sequences and can be modeled as gene conversion (Lapierre et al. 2016). Thus, the specifics of bacterial genome evolution require the formulation of appropriate population genetics models for prokaryotic genomes (Rocha 2018).

The prevalence of horizontal gene transfer (HGT) in bacteria, which can result in the transfer of genetic material between distantly related individuals (Arnold et al. 2022), might lead to the conviction that the biological species concept does not extend to prokaryotes. However, the rate of HR decreases with sequence divergence and breaks down at species boundaries, supporting HR as a species-forming force in prokaryotes (Bobay and Ochman 2017; Iranzo et al. 2019; Olm et al. 2020). Thus, the natural unit for estimating pangenomes and their shapes is a recombining population. The rates of the molecular mechanisms affecting the transfer of genetic material differ within and between populations. HR acts mostly within populations, but can nevertheless also occur between species at a lower rate (Doroghazi and Buckley 2010). HR can lead to gene conversion, where one or multiple variants are introduced into a homologous region, termed (allelic) recombination (Box 1). Additionally, HR can lead to the deletion or acquisition of genetic material in between the homologous flanking regions (Arnold et al. 2022) (Box 1). Next to HR, gene acquisition can be mediated by site-specific recombination of mobile genetic elements (MGEs), such as transposons, integrons, integrative conjugative elements, plasmids, and phages (Arnold et al. 2022). Most gene transfers due to

Box 1: Evolutionary and ecological processes that shape prokaryotic pangenomes

A. Molecular mechanisms that affect the gene content of an individual

Homologous recombination (HR): A molecular mechanism to replace DNA sequences with homologous flanking regions. The rate of HR decreases with sequence divergence. HR facilitates gene conversion, resulting in allelic recombination (see below), and also gene gain or loss when the flanking regions are homologous.

Site-specific recombination: Insertion of mobile genetic elements, for example, phages, at a particular site or motif (typically of length 4–12 bp) within the bacterial chromosome.

B. Fitness-independent population genetics processes that affect allele frequencies in prokaryotic populations

Mutation: A small-scale variant in a single nucleotide (substitution) or one or few nucleotides (insertion or deletion).

Allelic recombination (also termed recombination in population genetics models): The introduction of one or multiple mutations due to gene conversion by homologous recombination (see above) (Rocha 2018). The introduced mutation(s) depend on the population. Whereas allelic recombination reshuffles variants, the expected change of allele frequencies (ECAF) is zero.

C. Fitness-independent population genetics processes that affect gene frequencies in prokaryotic pangenomes

For each process, we state the **expected change of the gene frequency (ECGF)** of a single accessory gene that is present in **frequency x** in the population, if only this process is considered. Note that all processes occur independently of the fitness, but nevertheless add variation to the population (see also fig. 1).

Horizontal gene transfer (HGT): A general term for the transfer of genetic material between prokaryote individuals.

Acquiring gene transfer (AGT): A form of HGT denoting the transfer of genetic material between distinct prokaryotic populations. Notably, the acquisition depends on the gene frequency in the environment, where genes might already be present that enable niche-specific adaptation. If the gene is common in the environment, there is a constant influx of this gene into the bacterial population, which results in an ECGF of $q(1-x)$, where q is the gene gain rate via AGT. Furthermore, genes that are rare in the environment will occasionally add new genes to the population.

Spreading gene transfer (SGT): Another form of HGT, where the genetic material is transferred within a prokaryotic population. It acts on the genes that are already in the population and can reshuffle variants by creating new gene combinations. As the number of effective SGT events depends on the number of possible interactions between carriers and noncarriers, the frequency of a transferable gene will increase proportional to $gx(1-x)$, where g is the gene transfer rate. The effect of SGT is thus strongest for genes with intermediate frequencies.

Gene loss: Loss of the genes present in a population resulting in a linear decrease in gene frequencies. If gene loss occurs at rate r, the ECGF is given by $-rx$.

D. Fitness-independent population genetics processes that affect allele and gene frequencies

Genetic drift: Random changes in gene frequencies that arise due to the stochastic nature of inheritance and reproduction in a population. It can cause the frequency of an allele or gene to increase or decrease and can also cause loss or fixation of a gene in the population by chance. The ECAF and the ECGT due to genetic drift are 0, meaning that over time, the frequency of an allele or gene will fluctuate starting from its initial value in a random manner.

Migration: Movement of individuals between niches, which can decrease or increase the frequency of an allele or a gene, depending on its initial frequency in those niches. If individuals immigrate at rate m from a population with allele/gene frequency y, the ECAF/ECGF is $-m\ (x-y)$. ECGF for migration equals ECGF for gene loss if $y=0$ and $m=r$ and for AGT if $y=1$ and $m=q$.

E. Fitness-dependent population genetic process that affects bacterial diversity

Selection: The only evolutionary process that is adaptive. It results in a logistic growth of the selected genotypes. The spread of a variant under positive selection depends on the number of individuals that possess the variant and is modeled as an average increase in frequency that is proportional to $sx(1-x)$, where s is the selective benefit to possess the variant. Note that for pangenomes, the variant is a gene occurrence. The ECGF for selection is equivalent to the ECGF for SGT when $s=g$.

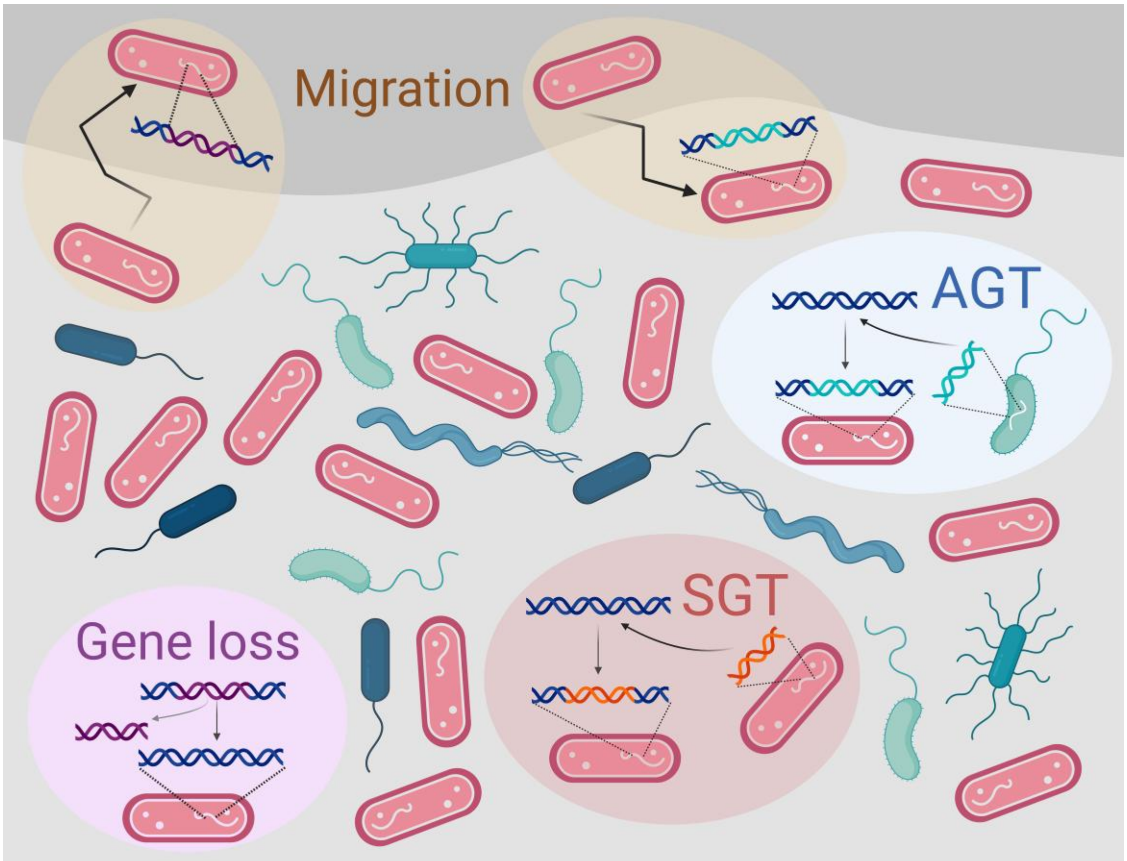

Fig. 1.—Scheme of the fitness-independent processes that shape prokaryotic pangenomes (Box 1C). Created with BioRender.com.

MGEs occur between closely related bacteria; for example, phages are often species- or even strain-specific. Nevertheless, MGEs can also mediate the acquisition from distantly related organisms in the environment (Khedkar et al. 2022), and especially the host breadth of plasmids is highly variable across plasmids groups, where some show wide host ranges beyond the species (Redondo-Salvo et al. 2020) (Box 2).

As gene transfer acts differently depending on how closely related the donor and recipient genomes are, we here propose to distinguish two different processes of gene gain when modeling prokaryotic pangenomes depending on where the donor gene comes from. First, gene gain from members of the same population results in the spread of genes within the population (termed here spreading gene transfer, SGT) and can be facilitated by HR and by MGEs. Second, gene transfer from distantly related organisms, leading to the acquisition of novel genes (termed here acquiring gene transfer, AGT), is mostly facilitated by MGEs. In the following, we argue that models for pangenome evolution must include a multitude of population genetics processes, including SGT and AGT (Box 1).

Major Population Genetic Processes Affecting Pangenomes Mimic Each Other When Considering Individual Genes

For point mutations, the nucleotide diversity is determined by the product of N_e and the mutation rate. As the mutation rate is often similar within species, nucleotide diversity is determined by variations in the demography and selective processes that affect N_e, which allows inference of these processes (Hartl 2020). Compared to mutation rates, gene transfer rates have a much larger variation between different genes and populations, resulting in inconclusive inferences of selection from the observed pangenomes (Bobay and Ochman 2018; Douglas and Shapiro 2021; Cummins et al. 2022).

Consequently, the existing population genetic theory for point mutations cannot be directly applied to model the gene gain and loss dynamics in pangenomes. When considering point mutations, a reasonable null model is the *infinite-sites model*, where mutations occur at a constant rate and each mutation hits a new position (Hartl 2020). Among the variants contributing to pangenomes—AGT, SGT, and gene loss—AGT is most comparable to mutation. Admittedly, both are substantially different biological

Box 2: Outstanding challenges when investigating SGT and AGT

Mobile genetic elements

The contribution of different classes of MGEs that are involved in gene transfer varies with respect to SGT and AGT. Depending on the host range, certain phages or plasmids may promote primarily SGT or AGT. Knowledge about the specificity of the involved MGEs could thus provide valuable prior information for model parametrization.

Homologous recombination (HR)

HR acts mostly within populations, but can nevertheless also occur between species at a lower rate, influencing rates of SGT and AGT. Additionally, bacterial population vary in their capability of natural transformation. A detailed understanding of these processes could support the model parametrization for a particular population.

Delimitation of prokaryotic populations

The distribution of empirical data and the mechanistic differences in gene transfer between closely or distantly related bacteria support that the rate of gene transfer depends on the similarity between donor and recipient sequences. Here, we propose to model this dependency by modeling the different processes of SGT and AGT with different rates. Nevertheless, there are no clear thresholds at which prokaryotic populations can be clearly delimited. The determination of SGT and AGT rates will depend on the chosen separation of populations, where the most natural delimitation of populations will depend on the natural SGT and AGT rates and is also influenced by the population structure and migration between subpopulation. Modeling SGT and AGT thus requires a joint consideration of population delimitation and migration.

Genome-wide linkage and co-transfer of genes

For eukaryotes, the degree of linkage between genetic loci is determined by their relative positions along the chromosome. This has allowed to development of concepts such as linkage disequilibrium and genetic recombination maps, which help to disentangle selective and recombining processes. For prokaryotes, clonal reproduction leads to genome-wide linkage but the flexible rearrangement of accessory genes makes the linkage structure non-monotonic and more complex. The effect of the global linkage structure on gene transfer processes that can co-transfer nearby genes, such as SGT, is not yet fully understood, and conversely, how these processes impact the linkage structure is still subject to further studies.

Selfish genes

Selection can occur at both the individual level and the gene level. At the individual level, the fitness is determined by the complete gene composition and the residing environment. At the gene level, selection also acts on the frequency of SGT, with genes located in more easily transferred regions, such as plasmids, having a higher chance of increasing in frequency. The uncertainty of whether one type of selection dominates in shaping bacterial pangenomes is a so far unsolved challenge.

Abundance of singleton and rare accessory genes

Both neutral and selective models struggle to explain the huge numbers of singleton and extremely rare accessory genes observed. Possible contributors for these cloud genes of rare accessory genes include 1) constant influx of slightly costly and almost neutral genes by AGT that have not yet been purged by selection, 2) genes that were more frequent previously and remain in the population in low frequency due to a balance between SGT and negative selection, 3) the existence of a huge number of ecological niches, and 4) sampling biases. The role of the cloud genes for bacterial adaptation and whether they are an irrelevant side effect of the unavoidable influx of external genetic material or a valuable resource that is maintained by SGT and selective forces remains so far unclear.

processes that lead to different evolutionary dynamics; for example, in contrast to mutations, genes are often gained or lost in sets (see below). Nonetheless, when considering the dynamics of a single site or gene, AGT introduces novel polymorphisms by bringing new genes into the population, just as mutation introduces novel polymorphisms by modifying gene sequences. In contrast to mutation, different gene categories have different acquisition rates via AGT, depending on the gene abundances in the environment. While, in principle, AGT can introduce a new rare gene at each event, genes that are enriched in the niche-specific gene pool provide a constant influx of the gene into the population (Polz et al. 2013). Selection and gene loss will counteract this influx and purge non-beneficial genes from the population, leading to a selection-gene-loss-gene-acquisition balance for the number of accessory genes in the population. If we were to know the gene acquisition frequency and gene losses would occur rarely, then the balance could be used to estimate the strength of selection. However, gene loss occurs frequently in prokaryotes. The balance of frequent gene loss and AGT of rare genes has been included in the neutral *infinitely many genes model* (Baumdicker et al. 2012). This model can reproduce the typical U-shape of gene frequency distributions in pangenomes, although relaxing some of its assumptions further improves the fit (Sela et al. 2021).

In contrast to AGT, the rate at which SGT spreads a gene within the population depends on the current frequency of the gene in the population. For rare genes, only a small fraction of donors can spread the gene, while an abundant gene will mostly be transferred to an individual that already possesses the gene. At intermediate gene frequencies though, interactions are more likely to occur between a gene carrier and a non-carrier leading to effective spread. Thus, despite being a non-selective process, SGT has the same effect on gene frequencies as positive selection on single genes (Box 1) (Tazzyman and Bonhoeffer 2013; Baumdicker and Pfaffelhuber 2014).

Consequently, the frequency distribution of an accessory gene that can be explained by positive selection will also be explainable solely by SGT at an appropriate rate, or any corresponding combination of selection strength and SGT rate. Furthermore, spread by SGT could counterbalance the frequency decrease due to weak selection, for example, caused by bioenergetic costs, resulting in the retention of these nearly neutral genes in the long run (van Dijk et al. 2020). Finally, the spread by SGT can also balance the loss of genes involved in public goods production in cheaters, thereby stabilizing cooperation in bacterial populations (Lee et al. 2022). Importantly, the SGT rate differs substantially between genes. Particularly, high gene-specific SGT rates enable the spread of selfish MGEs (Shapiro 2017; Douglas and Shapiro 2021; Haudiquet et al. 2022) (Box 2). Nevertheless, the persistence of mobile elements such as plasmids can be explained by further mechanisms besides SGT (Brockhurst and Harrison 2022).

Gene frequencies are also influenced by migration, which previously led to the argument that migration is necessary for pangenomes to emerge (McInerney et al. 2017). However, the population genetic definition for migration of individuals between niches with different gene frequencies (x and y in Box 1) allows a broad modeling flexibility and can thus mimic a variety of other mechanisms. First, if the gene is absent in the other ecological niche ($y = 0$ in Box 1), as, for example in the model of Niehus et al. (2015), immigrants reduce the gene frequency in the focal population, just as gene loss does. Consequently, the resulting mathematical term for neutral gene loss (Baumdicker et al. 2012) is identical to the term that describes the effect of migration from an ecological niche, where the gene is absent, into the focal population. Based on these models, it is impossible to distinguish the effects of migration and gene loss on the frequency of an individual gene in the population. Second, when the gene is present in all emigrating individuals of the other population ($y = 1$ in Box 1), the effect of migration on the focal population is equivalent to the effect of gene gain by AGT. Third, otherwise ($0 < y < 1$), migration pushes the gene frequency toward intermediate levels, similar to complex selective processes.

Complex Selection Processes Also Contribute to the Abundance of Accessory Genes

When individuals migrate between different niches, they are subject to variable selection pressure over time. Similarly, entire populations are subject to frequent changes in the direction of selection when the environment imposes fluctuating selection pressures. In contrast to diploids, standard population genetic models for haploids predict that fluctuating selection leads to fixation of the variant with the largest geometric mean fitness. However, extended models and empirical observations imply that fluctuating selection can also promote polymorphisms in haploids (Dean et al. 2017), suggesting that intermediate gene frequencies in bacterial pangenomes could also be the result of environmental fluctuations. Nevertheless, both spatial migration and temporal fluctuating selection will only result in the retention of genes that are beneficial in at least one environment.

Interestingly, 40% of the transfers between different *Bacillus subtilis* lineages, which include recombination as well as gene transfer, were found to be under positive selection under laboratory conditions (Power et al. 2021). In contrast, Conrad et al. (2022) found that 3.5% of the accessory genes of *Salinibacter ruber* became abundant when exposed to different salinity conditions, suggesting a selective benefit of these genes in that environment. While it might be unfeasible to measure all gene-specific selection strengths for different environmental conditions and genetic backgrounds, Acar Kirit et al. (2022) highlight that the distributions of fitness effects over genes follow a common shape across environments, which might be implemented in evolutionary models.

Another potential cause of pangenomes with an elevated number of genes of low and intermediate frequencies is negative frequency-dependent selection (NFDS), where the benefit provided by a gene decreases with its frequency in the population (Levin et al. 1988). For example, NFDS arises naturally for genes that provide immunity against coevolving phages or other mobile genetic elements (Corander et al. 2017), in line with the observed abundance of defense systems among accessory genes (Bernheim and Sorek 2019).

Considering Multiple Accessory Genes and Their Linkage Results in Genomic Patterns that Differ Between Different Population Genetic Processes

While processes such as SGT and selection are indistinguishable at the level of single gene dynamics, they are fundamentally different when considering the joint distribution of all accessory genes across individuals of a population. Alongside a gene under positive selection, all other accessory genes present in the same individuals will hitchhike and thus increase in frequency. In contrast, SGT typically leads at most to the co-transfer of a few neighboring genes. While this can also affect the frequencies of other accessory genes, the scope is more limited compared to hitchhiking. Similarly, while mimicking the effect of AGT and gene loss at the level of individual genes, migration affects all genes in a genome collectively.

When modeling the joint evolution of accessory genes, two opposing simplifying assumptions are often proposed. One assumes that gene-specific sweeps due to high SGT rates are common in bacterial populations (Shapiro and Polz 2014; Takeuchi et al. 2015). This suggests that accessory genes evolve independently and unlinked, which preserves the background diversity of the pangenome. The other assumes that the joint evolution of genes still depends on the clonal relationships or population structure (Horesh et al. 2021), suggesting that their evolution can be constrained to one phylogenetic tree. The relevance of the underlying clonal tree or population structure on gene linkage is currently debated (Sakoparnig et al. 2021; Preska Steinberg et al. 2022). However, associations of phenotype with gene presence–absence and gene co-occurrence analyses strongly depend on the linkage of individual genes (Saber and Shapiro 2020; Whelan et al. 2020). Co-occurrence of genes can be caused by vertical co-inheritance, horizontal co-transfer, co-selection, that is co-occurring selection pressures, or epistasis, that is when the fitness effect of one gene depends on the presence or absence of the other gene. Notably, the relative frequency of multi-gene transfers is higher between species compared to within species (Kloub et al. 2021), suggesting an important role for AGT in co-transfer. Thus, gene co-occurrences cannot be simply matched with the importance of selection, contrasting previous conclusions (McInerney 2022). Instead, measuring the fitness effect of genes in various backgrounds is essential when estimating epistatic effects, which has only been attempted for a few species so far (Acar Kirit et al. 2022; Rosconi et al. 2022).

To Understand the Processes That Shape Pangenomes, It Is Necessary to Consider Multiple Evolutionary Processes in Concert

Vast amounts of sequencing data from closely related strains already enabled detailed inferences on allelic recombination (Preska Steinberg et al. 2022). So far, pangenome properties were shown to vary between environments (Maistrenko et al. 2020). To disentangle the underlying relevant processes, we need three important components. First, to obtain an unbiased view of the pangenome and its gene frequencies, well-balanced sampling strategies of natural communities are necessary to mitigate sampling biases (Brockhurst et al. 2019; Liao et al. 2021). Second, to correctly assign sequences to homologous genes, we need accurate methods to annotate genomes and to reconstruct pangenomes. A multitude of tools exist, which apply different approaches and are thus differently affected by the various types of errors, such as misannotation, contamination, and clustering errors. The tool chosen can thus have a large influence on the quality of pangenome reconstruction (Ding et al. 2018; Tonkin-Hill et al. 2020).

Third and finally, in this perspective, we call for appropriate population genetic models and distinguishable summary statistics. Simulation tools are particularly relevant for complex models, where many previous studies relied on simulation tools tailored for the specific scenario considered (Niehus et al. 2015; van Dijk et al. 2020; Power et al. 2021). We have to keep in mind that specific modeling choices can limit the ability to discriminate different hypotheses and may lead to incorrect conclusions (Lapierre et al. 2016). To enable the comparison of different hypotheses, we need to incorporate all relevant processes into the same modeling framework and to standardize simulations, where considering the joint distribution of accessory genes across individuals of a population will be particularly relevant. Promising simulation tools to understand which processes potentially shaped pangenome diversity are emerging, such as the bacterial slimulator (Cury et al. 2022), stdpopsim (Adrion et al. 2020), BacGWASSim (Saber and Shapiro 2020), and Bacmeta (Sipola et al. 2018).

To conclude, it is an ill-posed problem to try to determine a single process that has shaped a particular prokaryotic pangenome. Pangenomes are formed by the concerted action of selective forces and fitness-independent processes acting within a population and of the selective advantage of (selfish) genes themselves. Additionally, to some degree, they might just be a glimpse of the vast amount of constantly inflowing environmental genes that have not yet been removed or fixed. All these processes should be considered jointly. So far, disregarded differences (e.g., between SGT and AGT) and equivalent effects (e.g., of SGT and positive selection) have hindered the generalization of findings and fueled the debate about the relative importance of neutral and selective evolution. The notion that evidence against selection is necessary to support a neutral view of pangenome evolution (McInerney 2022) is a logical pitfall that should be avoided. Models that include selective processes can generate the same diversity patterns as fitness-independent processes. Thus, if selective models fit the observed data, it does not necessarily imply that genes are under selective pressure. At the same time, when neutral models can generate the observed diversity in pangenomes, it does not follow inevitably that genes evolve neutrally.

Fortunately, when developing population genetics models for prokaryotic pangenomes, we can stand on the shoulders of giants and build upon the model complexity that we have achieved in diploids. To resolve the pangenome dilemma and disentangle the contribution of different processes based on the large number of whole genome sequences available for bacteria, we need a unifying modeling approach considering the relevant involved processes.

Acknowledgments

We thank Libera Lo Presti for helpful comments and proofreading. F.B. is funded by the Deutsche Forschungsgemeinschaft EXC 2064/1: Project number 390727645 and EXC 2124: Project number 390838134.

Data Availability

All data underlying this article are available in the article itself.

Literature Cited

Acar Kirit H, Bollback JP, Lagator M. 2022. The role of the environment in horizontal gene transfer. Mol Biol Evol. 39:msac220.

Adrion JR, et al. 2020. A community-maintained standard library of population genetic models coop. Elife 9:e54967.

Andreani NA, Hesse E, Vos M. 2017. Prokaryote genome fluidity is dependent on effective population size. ISME J. 11:1719–1721.

Arnold BJ, Huang I-T, Hanage WP. 2022. Horizontal gene transfer and adaptive evolution in bacteria. Nat Rev Microbiol. 20:206–218.

Baumdicker F, Hess WR, Pfaffelhuber P. 2012. The infinitely many genes model for the distributed genome of bacteria. Genome Biol Evol. 4:443–456.

Baumdicker F, Pfaffelhuber P. 2014. The infinitely many genes model with horizontal gene transfer. Electron J Probab. 19:1–27.

Bernheim A, Sorek R. 2019. The pan-immune system of bacteria: antiviral defence as a community resource. Nat Rev Microbiol. 18: 113–119.

Bobay L-M, Ochman H. 2017. Biological species are universal across Life's Domains. Genome Biol Evol. 9:491–501.

Bobay L-M, Ochman H. 2018. Factors driving effective population size and pan-genome evolution in bacteria. BMC Evol Biol. 18:153.

Brockhurst MA, et al. 2019. The ecology and evolution of pangenomes. Curr Biol. 29:R1094–R1103.

Brockhurst MA, Harrison E. 2022. Ecological and evolutionary solutions to the plasmid paradox. Trends Microbiol. 30:534–543.

Comeron JM. 2017. Background selection as null hypothesis in population genomics: insights and challenges from *Drosophila* studies. Philos Trans R Soc Lond B Biol Sci. 372:20160471.

Conrad RE, et al. 2022. Toward quantifying the adaptive role of bacterial pangenomes during environmental perturbations. ISME J. 16:1222–1234.

Corander J, et al. 2017. Frequency-dependent selection in vaccine-associated pneumococcal population dynamics. Nat Ecol Evol. 1:1950–1960.

Cummins EA, Hall RJ, Connor C, McInerney JO, McNally A. 2022. Distinct evolutionary trajectories in the *Escherichia coli* pangenome occur within sequence types. Microb Genom. 8:000903.

Cury J, Haller BC, Achaz G, Jay F. 2022. Simulation of bacterial populations with SLiM. Peer Community J. 2:e7.

Dean AM, Lehman C, Yi X. 2017. Fluctuating selection in the Moran. Genetics 205:1271–1283.

Ding W, Baumdicker F, Neher RA. 2018. panX: pan-genome analysis and exploration. Nucleic Acids Res. 46:e5.

Doroghazi JR, Buckley DH. 2010. Widespread homologous recombination within and between Streptomyces species. ISME J. 4: 1136–1143.

Douglas GM, Shapiro BJ. 2021. Genic selection within prokaryotic pangenomes. Genome Biol Evol. 13:evab234.

Hartl DL. 2020. A primer of population genetics and genomics. Oxford, UK: Oxford University Prfess.

Haudiquet M, de Sousa JM, Touchon M, Rocha EPC. 2022. Selfish, promiscuous and sometimes useful: how mobile genetic elements drive horizontal gene transfer in microbial populations. Philos Trans R Soc Lond B Biol Sci. 377:20210234.

Horesh G, et al. 2021. Different evolutionary trends form the twilight zone of the bacterial pan-genome. Microb Genom. 7:000670.

Iranzo J, Wolf YI, Koonin EV, Sela I. 2019. Gene gain and loss push prokaryotes beyond the homologous recombination barrier and accelerate genome sequence divergence. Nat Commun. 10:5376.

Jensen JD, et al. 2019. The importance of the Neutral Theory in 1968 and 50 years on: a response to Kern and Hahn 2018. Evolution 73: 111–114.

Johri P, Eyre-Walker A, Gutenkunst RN, Lohmueller KE, Jensen JD. 2022. On the prospect of achieving accurate joint estimation of selection with population history. Genome Biol Evol. 14:evac088.

Khedkar S, et al. 2022. Landscape of mobile genetic elements and their antibiotic resistance cargo in prokaryotic genomes. Nucleic Acids Res. 50:3155–3168.

Kimura S. 1983. The neutral theory of molecular evolution. Cambridge, UK: Cambridge University Press.

Kloub L, et al. 2021. Systematic detection of large-scale multigene horizontal transfer in prokaryotes. Mol Biol Evol. 38: 2639–2659.

Lapierre M, Blin C, Lambert A, Achaz G, Rocha EPC. 2016. The impact of selection, gene conversion, and biased sampling on the assessment of microbial demography. Mol Biol Evol. 33:1711–1725.

Lee IPA, Eldakar OT, Gogarten JP, Andam CP. 2022. Bacterial cooperation through horizontal gene transfer. Trends Ecol Evol. 37: 223–232.

Levin BR, et al. 1988. Frequency-dependent selection in bacterial populations. Philos Trans R Soc Lond B Biol Sci. 319:459–472.

Liao J, et al. 2021. Nationwide genomic atlas of soil-dwelling *Listeria* reveals effects of selection and population ecology on pangenome evolution. Nat Microbiol. 6:1021–1030.

Lynch M, Marinov GK. 2015. The bioenergetic costs of a gene. Proc Natl Acad Sci U S A. 112:15690–15695.

Maistrenko OM, et al. 2020. Disentangling the impact of environmental and phylogenetic constraints on prokaryotic within-species diversity. ISME J. 14:1247–1259.

McInerney JO. 2022. Prokaryotic pangenomes act as evolving ecosystems. Mol Biol Evol. 40:msac232.

McInerney JO, McNally A, O'Connell MJ. 2017. Why prokaryotes have pangenomes. Nat Microbiol. 2:17040.

Mira A, Ochman H, Moran NA. 2001. Deletional bias and the evolution of bacterial genomes. Trends Genet. 17:589–596.

Niehus R, Mitri S, Fletcher AG, Foster KR. 2015. Migration and horizontal gene transfer divide microbial genomes into multiple niches. Nat Commun. 6:8924.

Olm MR, et al. 2020. Consistent metagenome-derived metrics verify and delineate bacterial species boundaries. mSystems. 5:e00731-19.

Polz MF, Alm EJ, Hanage WP. 2013. Horizontal gene transfer and the evolution of bacterial and archaeal population structure. Trends Genet. 29:170–175.

Power JJ, et al. 2021. Adaptive evolution of hybrid bacteria by horizontal gene transfer. Proc Natl Acad Sci U S A. 118:e2007873118.

Preska Steinberg A, Lin M, Kussell E. 2022. Core genes can have higher recombination rates than accessory genes within global microbial populations. Elife 11:e78533.

Redondo-Salvo S, et al. 2020. Pathways for horizontal gene transfer in bacteria revealed by a global map of their plasmids. Nat Commun. 11:3602.

Rocha EPC. 2018. Neutral theory, microbial practice: challenges in bacterial population genetics. Mol Biol Evol. 35:1338–1347.

Rosconi F, et al. 2022. A bacterial pan-genome makes gene essentiality strain-dependent and evolvable. Nat Microbiol. 7:1580–1592.

Saber MM, Shapiro BJ. 2020. Benchmarking bacterial genome-wide association study methods using simulated genomes and phenotypes. Microb Genom. 6:e000337.

Sakoparnig T, Field C, van Nimwegen E. 2021. Whole genome phylogenies reflect the distributions of recombination rates for many bacterial species. eLife 10:e65366.

Sela I, Wolf YI, Koonin EV. 2021. Assessment of assumptions underlying models of prokaryotic pangenome evolution. BMC Biol. 19:27.

Shapiro BJ. 2017. The population genetics of pangenomes. Nat Microbiol. 2:1574.

Shapiro BJ, Polz MF. 2014. Ordering microbial diversity into ecologically and genetically cohesive units. Trends Microbiol. 22: 235–247.

Sipola A, Marttinen P, Corander J, Berger B. 2018. Bacmeta: simulator for genomic evolution in bacterial metapopulations. Bioinformatics 34:2308–2310.

Takeuchi N, Cordero OX, Koonin EV, Kaneko K. 2015. Gene-specific selective sweeps in bacteria and archaea caused by negative frequency-dependent selection. BMC Biol. 13:20.

Tazzyman SJ, Bonhoeffer S. 2013. Fixation probability of mobile genetic elements such as plasmids. Theor Popul Biol. 90:49–55.

Tonkin-Hill G, et al. 2020. Producing polished prokaryotic pangenomes with the Panaroo pipeline. Genome Biol. 21:180.

van Dijk B, Hogeweg P, Doekes HM, Takeuchi N. 2020. Slightly beneficial genes are retained by bacteria evolving DNA uptake despite selfish elements. Elife 9:e56801.

Vos M, Eyre-Walker A. 2017. Are pangenomes adaptive or not? Nat Microbiol. 2:1576.

Whelan FJ, Rusilowicz M, McInerney JO. 2020. Coinfinder: detecting significant associations and dissociations in pangenomes. Microb Genom. 6:000338.

Associate editor: Laura Eme