Article:
Brockhurst, M.A., Harrison, E., Hall, J.P.J. et al. (3 more authors) (2019) The ecology and evolution of pangenomes. Current Biology, 29 (20). R1094-R1103. ISSN 0960-9822

https://doi.org/10.1016/j.cub.2019.08.012

THE ECOLOGY AND EVOLUTION OF PANGENOMES

Michael A Brockhurst[1*], Ellie Harrison[1], James PJ Hall[2], Thomas Richards[3], Alan McNally[4], Craig MacLean[5]

* Corresponding Author

1. Department of Animal and Plant Sciences, University of Sheffield, Sheffield UK
2. Institute for Integrative Biology, University of Liverpool, Liverpool UK
3. Biosciences, University of Exeter, Exeter UK
4. Institute of Microbiology and Infection, University of Birmingham, Birmingham UK
5. Department of Zoology, University of Oxford, Oxford UK

Abstract

The pangenome is all the genes present in a species and can be subdivided into the accessory genome, present in only some of the genomes, and the core genome, present in all the genomes. Pangenomes arise due to gene gain by genomes from other species through horizontal gene transfer and differential gene loss among genomes. Our current view of pangenome variation is phenomenological and incomplete. We outline the mechanistic, ecological and evolutionary drivers of and barriers to horizontal gene transfer that are likely to structure pangenomes, highlighting the key role of conflict between the host chromosome(s) and the mobile genetic elements that mediate gene exchange. We identify shortcomings in our current models of pangenome evolution and suggest directions for future research to allow a more complete understanding of how and why pangenomes evolve.

The pangenome concept

The pangenome describes all the genes present in a species and can be subdivided into those shared by all members of a species—the core genes—and those present in only some members of a species—the accessory genes [1] (Figure 1). Although a pangenome can be defined for other taxonomic units (e.g., an ecotype or phylum), we focus here on the single species level since this is the most commonly used meaning. The pangenome concept emerged from early comparative studies of bacterial genomes. Comparison of a pathogenic

Escherichia coli O157 strain with its non-pathogenic relative *E. coli* K12, showed substantial gene gain in the O157 genome [2]. Shortly afterwards, a three-way comparison of these two genomes with that of another pathogenic *E. coli* genome, showed that less than 40% of protein coding sequences were shared between all three strains despite all being members of the *E. coli* species [3], which has proven to have an exceptionally broad pangenome. Even in these early pangenome studies it was evident that the variation among genomes within a species is often attributable to horizontal gene transfer (HGT) events. For instance, the difference between the *E. coli* strains K12 and O157 genomes is largely due to the acquisition of several large pathogenicity islands by O157 [2]. This variation is part of a wider pattern of variation in pathogenicity islands seen across *E. coli*, where differential distribution in these genomic regions is responsible for the classical nomenclature of *E. coli* pathotypes [4]. These range from chromosomally integrated pathogenicity islands and prophages to independently replicating plasmids. The advent of next-generation sequencing brought with it an acceleration in the generation of bacterial genome sequence data, revealing that the size of the pangenome varies widely among taxa. These studies reveal an overall negative relationship between pangenome size and the proportion of core genes: “open” pangenomes are larger in size, have a smaller proportion of core genes, and higher rates of gene gain by HGT, whereas “closed” pangenomes are smaller in size, have a larger proportion of core genes, and lower rates of gene gain by HGT (Figure 1) [5]. The concept of a pangenome in eukaryotes is debated [6, 7], but the available genomic data suggests that the concept is sound, although the extent of the accessory genome and the processes that drive the evolution of pangenome content are in many ways different in eukaryotes compared to prokaryotes (Box 1).

The current challenge is to move beyond this phenomenological description of pangenomes to forge an understanding of the mechanisms and processes that determine their structure. A genome sequence is a snapshot of a strain in time. Some of the genes and mutations in that snapshot share a long history and are destined to remain associated, while other members are transient: recent acquisitions in the process of leaving. How do we distinguish between these categories? If a genome is a family photograph, how do we distinguish family members from the photobombers? A starting point is to understand the processes and mechanisms that promote or prevent gene gain and loss, and thereby shape the content of the pangenome. Gene gain by a lineage in the context of the pangenome can be conceptually separated into two distinct processes, operating on different timescales and affected by different environmental drivers. The first describes the specific gene acquisition event, which occurs at the level of individual cells and is effectively instantaneous, while the second represents the stable assimilation of acquired genes within populations or their non-random elimination from a lineage, and is on-going, with effects emerging over a longer period and in different ways in different environments. In this review, we first outline the molecular, ecological and evolutionary drivers of gene gain and loss which mediate changes in the composition of the pangenome, and then discuss how evolutionary theory can be applied to understand the structure of pangenomes.

Drivers and barriers of gene gain and loss

Gene acquisition introduces variation, and thus provides the raw material upon which selection can subsequently act [8]. Various mechanisms actively facilitate the movement of genetic material across membranes. These are particularly well-described in prokaryotes but there is evidence that equivalent mechanisms may exist in model eukaryotes such as yeast (see Box 1). In recent decades, the canonical processes — conjugation, transduction, and transformation — have been joined by additional phenomena, including nanotubes [9] and vesicles [10] that can facilitate nucleotide exchange. These varied mechanisms of gene exchange offer the potential for gene acquisition, but the likelihood of its occurrence depends on a range of ecological, mechanistic and evolutionary factors, explored in this section (summarised in Figure 2).

Ecological opportunity for HGT

The proximal environmental triggers activating expression of gene exchange machinery vary between systems and with different species, but some common themes can be identified. One of these is stress. For example, the SOS response to DNA damage, triggered by some antibiotics, reactive oxygen, and UV radiation, activates transfer of the Vibrio cholerae STX element [11], causes integron rearrangement [12], and activates integrated bacteriophage [13]. Transposons in *E. coli* become active under nutritional stress [14], plasmid conjugation rates are increased in response to host inflammation in mammalian gut [15], and starvation conditions activate natural competence [16]. However, different stress responses can have divergent effects in different species [17], and donors, recipients, and mobile genetic elements may each have their own cues. For example, some mobile genetic elements, such as the pheromone-inducible conjugative plasmids of *Enterococcus*, have evolved mechanisms to detect the presence of recipients [18], and transformation is induced by quorum sensing and by specific nutrients in some species of *Vibrio* [19].

Ecology appears to be a principal determinant of gene-sharing [20], suggesting that the transfer of genes is to some extent limited by ecological opportunity and occupancy of shared habitats. Several gene transfer mechanisms including conjugation and nanotubes require close physical proximity and thus HGT is probabilistically likely to be most efficient between immediate neighbours [21]. Consequently, the size of the gene pool from which a species can draw will be dependent on the diversity of environments they occupy as well as the community diversity these contain. Correspondingly, networks of gene sharing have shown that co-occurrence of species in a habitat increases the probability of gene sharing [22-25]. Niche specialists likely to exist in stable environments with very low diversity, such as endosymbionts [24], have more closed pan-genomes than those that exist in diverse communities and more variable environments.

Among symbionts and pathogens with low rates of gene gain through HGT, variation in gene loss among lineages can be the primary cause of diversity among clonal lineages, and can lead to large phenotypic differences [26]. Whereas gene loss can be positively selected in large populations with efficient selection, in intracellular symbionts and pathogens with low effective population size, gene loss is more likely to be a result of relaxed selection and drift [27]. How the balance of gene gain and loss contributes to the formation of a pangenome is well-illustrated by *Yersinia enterocolitica*. The species is composed of five phylogenetically distinct groups, four of which are pathogenic to humans and have emerged from a non-pathogenic ancestor, driven by a single acquisition of a large virulence plasmid [28]. Following plasmid acquisition, the splits between the four pathogenic groups are delineated at a pangenome level by differential losses of genes present in the ancestor, alongside HGTs leading to switches in serotype [29].

Mechanistic drivers and barriers of HGT

Once acquired there are significant barriers to the maintenance of novel genetic material which shape the patterns of gene sharing among species. Newly acquired DNA must replicate to ensure it is passed to daughter cells, either by carrying with it replication machinery compatible with that of the host (in the case of plasmids) or by integrating into a resident replicon (e.g. a chromosome or already-present plasmid). Integration can occur through general recipient-encoded processes such as homologous recombination which is dependent on regions of sequence homology flanking the heterologous gene [30, 31] or by the activity of entities such as transposons, integrons, and insertion sequences, which can facilitate capture of incoming DNA (e.g., [32]).

Genes must also be transferable and able to function in the host in order to have a phenotypic effect visible to selection [33], which is dependent on recognition of promoters allowing for gene expression [34], and comparable GC content, codon usage and compatible genetic codes allowing for efficient translation [35], and in the case of DNA transfer between eukaryotic genomes effective splicing of introns. Newly acquired genes evolve faster than older genes in the same genome, potentially because of adaptation to their new genomic context [36, 37]. As a general principle, many of these processes become more challenging across larger genetic distances [38]. Correspondingly gene sharing has been shown to be most common between closer phylogenetic relatives [25], which enhances both the likelihood of the transfer event and the compatibility of genes between donor and recipient.

Mechanistic limitations are also likely to define the types of genes that are more readily shared, and therefore more likely to contribute to the accessory genome. Incoming DNA can disrupt cellular processes leading to severe fitness costs, and these genes are likely to be rapidly lost from the population by purifying selection. Genes encoding core cellular functions, such as those associated with transcription and translation, can be highly toxic when expressed in foreign hosts [34, 39] and are poorly represented among horizontally transferred genes [40, 41]. This strong incompatibility may be due to disruption of or failure to maintain the large number of protein-protein interactions that the protein must engage in to properly function. Genes embedded within more complex interaction networks are therefore more disruptive and less likely to maintain the necessary functional interaction network when transferred, a phenomenon termed the complexity hypothesis [42, 43]. Mobile genetic elements (MGEs) themselves are often associated with significant fitness costs that are caused by a range of factors, including the biosynthetic cost of maintaining and expressing additional DNA, toxic gene products, and epistasis between chromosomal and MGE-encoded genes [44]. This disruptive effect of HGT is not surprising from an evolutionary perspective: HGT brings together genes that have different evolutionary histories, and there is no a priori reason to expect that these genes should function together harmoniously [45].

Evolutionary conflict and collaboration in the pangenome

Many of the mechanisms for horizontal gene transfer are encoded by infectious MGEs such as viruses, plasmids, and transposable elements. Therefore, pangenomes are composites of the host chromosome(s) together with MGEs that may be shared with other species. MGEs encode accessory genes that may represent adaptive additions to the pangenome (e.g. by providing a new ecological function or access to an otherwise inaccessible niche), but also encode genes for selfish MGE-directed functions such as replication and transmission, as well as many genes of unknown function. As semi-autonomous evolving entities we should expect MGEs to maximise their own fitness through both vertical and horizontal transmission [46]. Encoding beneficial accessory genes can increase MGE fitness through enhanced vertical transmission as positive selection drives clonal expansion [47]. However, being beneficial is not necessary for MGE success. Many environmental plasmids do not encode any obvious accessory genes [48] and are therefore likely to be genetic parasites. Experimental studies show that high rates of horizontal transmission through conjugation can maintain costly resistance plasmids in the absence of positive selection [47, 49, 50], and non-beneficial plasmids can invade biofilm populations [51, 52]. Indeed, experiments with antibiotic resistance and mercury detoxification plasmids have shown that positive selection for these functions can limit their horizontal transfer by reducing the availability of recipient cells [47, 53]. Although, in the long run, purely infectious elements would be expected to become increasingly efficient parasites by shedding their accessory genes, mobile genetic elements that persist through horizontal transmission are likely to be especially prone to mediating gene exchange [54]. Higher rates of horizontal transmission expose these MGEs to a wider diversity of genomic environments, offering greater opportunity for other MGEs (e.g., transposons) to integrate and hitch a ride. This inherent nestedness of pangenomes means that potentially conflicting selective pressures may operate at different levels of complexity (e.g., at the level of the gene, MGE, genome, population, and species etc.).

The predominance of gene exchange mediated by MGEs means that this form of gene sharing is, at least partially, constrained by MGE host range. Phages are believed to have relatively narrow host ranges, which are often limited to within a species or genus [55, 56]. Plasmid host ranges can be broader, and are dependent on the diversity of replication genes required for stable maintenance in different host taxa [57]. Correspondingly, plasmids appear to be more important mediators of gene exchange across larger genetic distances [58]. However, interactions between MGEs allow smaller, simpler elements to escape these restrictions.

Transposons for example, which are themselves unable to transfer between cells, can hitch a ride on a conjugative plasmid, as has been observed for plasmid-encoded antibiotic resistances in hospital outbreaks of Enterobacteriaceae [59, 60]. Further transfer of transposons between plasmids with different host ranges then expands the range of potential hosts accessible to these transposon-encoded genes. Plasmids too can be composite mosaics of other elements, including other plasmids, broadening the range of hosts in which they can replicate, while transposons can become nested within one another, increasing opportunities for spread [61, 62]. A consequence of the self-interested activity of MGEs for genome evolution is that selfish genes encoding MGE-related functions spread between lineages alongside the MGE-encoded accessory functions that enhance host fitness or niche adaptation. Indeed, plasmid, phage, and transposon-encoded functions are usually highly represented in the pangenome and in comparative studies of horizontal gene transfer [5, 63].

Because they can replicate by both vertical and horizontal transmission, MGEs can have fitness interests that do not necessarily align with those of other parts of the (vertically-inherited) genome. These 'divided loyalties' manifest in the fitness costs associated with MGE acquisition and horizontal transmission, and result in intragenomic conflict. For example, while conjugation provides an efficient mechanism for plasmids to transfer between bacteria, the expression of conjugative machinery imposes a biosynthetic fitness cost on the donor cell [64], and leaves the donor cell open to predation by pilus-targeting phage [65]. Resolution of host-MGE conflict frequently requires compensatory mutation(s) to the MGE or the chromosome to reduce the fitness costs of the newly acquired genes [46], which is promoted by positive selection for MGE-encoded functions since this increases the population size and mutation supply for MGE-carriers [66, 67]. Diverse compensatory mechanisms have been identified to stabilise plasmids, but two common routes are mutations affecting host gene regulatory networks [68, 69] or plasmid replication [45, 70]. By stabilising MGEs within bacterial lineages, compensatory evolution can set the stage for more extensive coevolution between the MGE and chromosome, driving reciprocal adaptations and counter-adaptations [46]. For example, bacteria-plasmid coevolution rapidly led to the emergence of co-dependence of chromosomal and plasmid replicons under antibiotic selection, together providing high-level resistance but separately providing inadequate levels of resistance to persist in the environment they evolved in [71, 72]. Compensation and coevolution can, in turn, drive the complete domestication of MGEs and their integration into a more exclusively vertical mode of replication. In practice, domestication involves downregulation, inactivation, or loss of the machinery involved in horizontal transmission [73, 74]. For example, bacterial genomes contain numerous prophages, some of which are incapable of horizontal transmission and now serve their bacterial hosts as anti-competitor toxins [75]. Alternatively, recombination can relocate mobile genes to less-mobile parts of the genome, e.g. chromosomal capture of resistance genes from plasmids, a process rapid enough to be readily observable in the laboratory [50, 69, 76]. In so doing, the signatures of gene acquisition are gradually lost from the genome sequence, potentially explaining why many accessory genes originally transferred by an MGE are no longer obviously associated with MGEs.

Resisting HGT

Due to the potential for conflict between MGEs and the host chromosome, immunity systems which actively target incoming foreign DNA are widespread across eukaryotes and prokaryotes. Systems exist in both eukaryotes (e.g. RNAi [77]) and prokaryotes (e.g. H-NS [78]) to silence gene expression from foreign DNA. In prokaryotes CRISPR-Cas systems and restriction-modification (R-M) systems target novel DNA for degradation, and can be an effective defence against MGEs, potentially reducing HGT [79, 80]. A comparative analysis of 79 prokaryote genomes show that R-M systems structure gene sharing by favouring exchanges between genomes with similar R-M systems [81]. The relationship between HGT and CRISPR-Cas systems appears more complex: There are well-described cases where CRISPR-Cas systems are negatively associated with MGE carriage within a species [82], but CRISPR-Cas can also promote HGT in some cases [83]. Type-III CRISPR-Cas systems target actively transcribed DNA via spacers derived from RNA transcripts [84] and may therefore be more effective against phages and plasmids than DNA acquired by transformation [85]. Over broader taxonomic scales, however, the correlation between CRISPR-Cas systems and the rate of HGT is less clear and deserves further study [86, 87]. It is likely that additional mechanisms for resisting gene acquisition will continue to be discovered [88]. Resistance mechanisms protecting cells against incoming DNA can also be encoded by MGEs themselves, highlighting how conflict between MGE could act to limit HGT. Both plasmids and phages defend their host cells against super-infection though self-exclusion mechanisms [89, 90] and can encode their own CRISPR-Cas systems with spacer sequences targeting other MGEs [91].

How and why do pangenomes evolve?

The next step is to synthesise these varied drivers of gene gain and loss into a general theory of pangenome evolution to answer the question: what structures the pangenome? On the one hand, it is conceivable that the pangenome is dominated by adaptive gene gain and loss, such that the pangenome is effectively a record of the responses to the myriad selection pressures that a species faces. At the other extreme, it is possible that the pangenome exists because selection is unable to prevent the spread of mildly deleterious gene acquisitions and deletions, and/or that these occur primarily due to the self-interest of MGEs. The key to distinguishing between these competing models of the pangenome is to disentangle how gene acquisition and loss, genetic drift, population subdivision and selection interact to shape the pangenome.

Population genetic approaches to analysing the pangenome

Evolutionary biologists have developed a mature body of population genetic theory to understand how mutation, selection and genetic drift interact to shape patterns of genetic variation [92]. A key insight from population genetic theory is that effective population size

(N_e) shapes patterns of molecular evolution by modulating the efficacy of natural selection relative to genetic drift [93]. In species with a low N_e, selection is weak relative to the genetic drift and evolution is dominated by the stochastic spread of weakly deleterious mutations. In contrast, selection prevents the spread of weakly deleterious mutations and drives selective sweeps of beneficial mutations in species with high N_e. Like spontaneous mutation, both gene acquisition [38, 44, 94, 95] and loss [96-98] tend to reduce fitness. Therefore, selection should shape patterns of gene gain and loss in species with high N_e, whereas the composition of the pangenome in species with low N_e will be shaped by underlying rates of gene gain and loss.

Genome size increases with N_e across a wide range of bacteria [99, 100], and this correlation provides a good starting point for applying population genetic approaches to understand the pangenome. In part, this correlation is driven by the inability of natural selection to prevent the spread of weakly deleterious mutations in species with low N_e [101], such as endosymbiotic bacteria [102] and intracellular pathogens [103]. Many genes in bacterial genomes only provide a fitness benefit under very specific environmental conditions [96], and effective selection for marginally beneficial genes acquired by HGT in species with high N_e is also likely to contribute to the positive correlation between N_e and genome size. Simply put, because species with large N_e are likely to occupy wider environment profiles, they are also likely to be under a wider diversity of environmental conditions driving selection for gene diversity and therefore larger genome sizes (Figure 1). As such species with high N_e also have large pangenomes [5, 100], and McInerney et al. [5] argue that this correlation is evidence that the pangenome is adaptive. The concept of population structure is key to this argument: in species with low levels of population structure, adaptive gene acquisition and loss events will sweep to fixation, and these will therefore not contribute to the pangenome. Population subdivision provides the opportunity for selection to contribute to increasing the pangenome size of a species because selective sweeps of locally adaptive gene gain and loss events will affect the accessory gene complement and thus pangenome size [104]. The point at which ecologically and genetically distinct subpopulations (or ecotypes) become sufficiently diverged to be considered multiple, different species each with their own pangenome is contentious [33, 105].

Other studies using population genetics have questioned the role of selection in shaping the pangenome. Comparing levels of synonymous nucleotide diversity, a surrogate measure of N_e, with a measure pangenome fluidity showed a positive correlation between N_e and pangenome fluidity, that could arise because genetic drift leads to the loss of effectively neutral accessory genes in species with low N_e [106]. Further support for this idea comes from comparing the observed distribution of gene frequencies in the pangenome with an expected distribution generated by a neutral model. This approach, inspired by the infinite alleles model, assumes that bacteria gain genes from an infinite pool of horizontally transferred genes and subsequently lose these genes through drift [107, 108]. Accessory genes show a distribution that is close to the expectations of a neutral model for widely distributed marine bacteria, but with deviations that are consistent with selection shaping the pangenome [108]. It is unclear, however, that currently available genomic data provide the necessary breadth and depth of ecological sampling to adequately test these models.

The limits of a population genetic approach

Population genetics theory provides some simple guiding principles for understanding the pangenome, but there are also potential difficulties with applying these models to understand the pangenome [109]. For example, classical population genetic tests for selection rely on comparing observed patterns of genetic polymorphisms and divergence with expected patterns from a neutral model where evolution is driven by mutation and drift, but not selection. Neutral models in population genetics assume that mutations at different sites in the genome are not linked. This is a justifiable assumption in eukaryotic species with obligate sexual reproduction, but the pangenome changes through the gain and loss of blocks of genes, for example because they are all encoded on a MGE. An important consequence of this is that strong selection for one gene (e.g. an antibiotic resistance gene) can lead to the spread of linked mildly deleterious genes by co-selection, if there is a net fitness benefit of the MGE. Similarly, genes that are linked to addiction systems, such as toxin-antitoxin systems, can be maintained in populations by the toxic effects of MGE loss. In a broader perspective, the strong linkage disequilibrium observed in clonal bacterial species means that there might be no effectively neutral variation [109].

A second important difficulty is that population genetic models ignore the evolutionary conflicts of interest that can occur between MGE-encoded accessory genes and chromosomal core genes in the same genome where selection at the MGE and chromosomal levels are not aligned. A key concept from evolutionary ecology is that trade-offs exist between the efficacy of vertical and horizontal transmission [110], preventing the evolution of elements that are to provide a big benefit to their host and transfer efficiently between hosts. Trade-offs may also limit the ability of MGEs to maximize the fitness benefit that they provide to different hosts, further limiting the benefits that hosts gain from acquiring MGEs [72]. All else being equal, we would therefore expect that MGEs with high mobility, such as broad-host range conjugative plasmids and lysogenic phage, to impose greater fitness costs than genetic elements with a low mobility, such as non-transmissible plasmids and defective prophage. This logic is somewhat counter-intuitive, because many of the pangenome accessory genes with the clearest ecological functions, such as antibiotic resistance genes, are often found on MGEs with high mobility [111-113]. These potentially adaptive genes may be rare ‘rubies in the rubbish’ from the perspective of their bacterial hosts [8], with the rest of the linked genes being either merely useless or else functioning solely to promote their own replication and transmission at the host’s expense.

Perspective

Short-read sequencing technologies have produced a rapid accumulation of sequence data, revealing the ubiquity and extent of pangenomes, especially in prokaryotes. At present, however, we lack a unified theory to understand the forces structuring pangenomes, and this will probably require the development of new theory that links together concepts from evolutionary ecology and population genetics. To achieve this, there are some important obstacles that need to be overcome:

- Defining the concept of pangenome adaptation: Adaptation is the "process of optimisation of the phenotype under the action of natural selection" [114]. As a pangenome emerges as an analytical result from comparing multiple genomes, we must take care when specifying what adaptation means in this context, i.e. who or what is being optimised. While a pangenome *can* contain adaptive genes that are transferred between species, the pangenome does not evolve *for the purposes of* maintaining a pool of niche-adaptive genes. Instead, its contents are defined by selection occurring at lower organisational levels: the individual bacterial lineage that has acquired locally-beneficial genes, and the persistent MGE. Neither does a broadly adaptive pangenome imply that the accessory genes in a given genome are beneficial to that strain. Recent migration or gene acquisition can result in a strain carrying neutral or deleterious genes which have not yet been lost [115]. Finally, if the pangenome is defined as the sum-total of all genes in a species, improved sequencing resolution will increasingly capture transient events which are unlikely to be adaptive, inflating the size of the pangenome but diluting the signal of adaptation. Enhanced biological insight into the gene function, as well as bioinformatic tools that help us distinguish between transient associations and longer-term partnerships, will guard us from incorrectly inferring adaptation in such instances.
- Measuring the rates of HGT in nature: The rate of horizontal gene transfer is key to both the population genetic and eco-evolutionary perspectives on the pangenome, but our knowledge of rate of HGT in the wild remains very limited. It might be possible to measure these rate by using statistical methods to infer rates of HGT from genomic data, and experimental methods that allow the spread of genes to be measured under natural communities in real time using for example microcosm experiments [54, 116].

- Sampling genomes at ecologically-relevant scales: Microbial genomes are being sequenced at an incredible rate, but it is very challenging to understand sequence data in a population genetics context, there are often huge sampling biases in microbial sequence datasets (intensive sampling of clinical outbreaks is the most extreme example). Given the vast population size of microbes, we will only ever be able to achieve very sparse sampling of microbial genomes, even with the most ambitious sequencing projects. We therefore need to develop approaches to identify and sample ecologically coherent microbial populations [113] or ecotypes [33]. For example, it is clear that some microbial populations are structured at an incredibly fine scale, such as individual particles of detritus [117], and this structuring can play a key role in the evolution of the pangenome [104]. Comparing a small number of bacterial genomes sampled from many niches is likely to produce an abundance of rare accessory genes, but these could either represent adaptive accessory genes that are locally abundant but globally rare, or deleterious accessory genes that are both locally and globally rare. One key technological development that may help with this problem is to move from sequencing the genomes of bacterial isolates to single-cell sequencing of bacteria from environmental samples.
- Developing eco-evolutionary models of pangenome evolution: The neutral theory of molecular evolution has been so useful in revealing the action of natural selection because it makes quantitative and falsifiable predictions that be tested by comparing datasets. Given the complexity of forces shaping the pangenome it may be necessary to look outside genetics for potential approaches: Pangenomes share many characteristics with metacommunities, most notably the idea that entities (genes or species) are sampled from a pool to form discrete sets (genomes or communities) that share biological cohesiveness (pangenome or metacommunity). Metacommunity ecology has a well-developed body of theory to understand how communities are assembled and structured [118], which may help to unravel the processes causing the structure of pangenomes.

BOX 1: Do eukaryotes have pangenomes? The existence of pangenomes in eukaryotes is debated [6, 7]. What is evident is that, unlike the situation in prokaryotes, genome evolution in eukaryotes is dominated by processes other than HGT, including sexual recombination and gene duplication [119] often combined with domain reshuffling [120]. Nevertheless, HGT can and does occur: for example, *Saccharomyces* undergoes transformation under starvation conditions [121] and can receive DNA by conjugation from bacteria [122], although HGT from prokaryotes contributes less than 0.5% of the gene repertoire of *Saccharomyces* (reviewed in [123]). Additionally, a range of other mechanisms introduce genetic material into eukaryotic cytoplasm offering the potential for HGT, including: viral vectors [124], integration of viral fragments [125], RNA exchange [126], trophic interactions through phagocytosis of prey cells [127], and anastomosis of cell structures [123, 128]. The role of HGT in accessory genome variation is unclear, but likely to be less important than in prokaryotes and a relatively minor contributor compared to other factors like strain level duplication [129] and differential gene loss. Pangenome studies in eukaryotes are challenging due to their more complex genome architectures and a lack of replete genome-level sampling. Analyses of model fungi suggest core genome fractions of between 80-90% [129], whilst in the marine alga *Emiliania huxleyi*, 17% of genes present in the assembled genome of the model strain CCMP1516 were absent in four other strains, indicating a putative accessory genome [130]. Consistent with the complexity of eukaryotic genome architecture, distinct dispensable or supernumerary chromosomes systems are observed in some fungi that show signs of HGT derivation, operate to carry an accessory genome, and define the niche and host range of the recipient lineage [131-133]. Therefore, while the existing studies suggest that the pangenome concept is well-founded for eukaryotic microbes, the extent of accessory genome variation is likely to be far lower than in prokaryotes: ~10-15% of genes in eukaryotes compared to up to ~65% in some prokaryotes.

Figure 1: The pangenome concept. Pangenomes vary extensively in size and the proportion of core versus accessory gene content. It is likely that species with large, open pangenomes occupy more varied niches and more complex communities, and have larger effective population sizes compared to species with smaller pangenomes.

Figure 2: The drivers and barriers of horizontal gene transfer. Horizontal gene transfer is likely to be affected by a wide range of ecological, evolutionary and mechanistic factors, which will in turn determine the degree of pangenome fluidity observed in a species.

Acknowledgements

This work was funded by grants from BBSRC (BB/R014884/1 to MAB, JPJH, EH; BB/R006253/1 to MAB; BB/R006261/1 to AM), NERC (NE/P017584/1 to EH; NE/R008825/1 to MAB, JPJH, EH; NE/S000771/1 to MAB) and the Wellcome Trust (106918/Z/15/Z to CM). The authors thank Maureen O’Malley and two anonymous reviewers for constructive comments on the manuscript.

Cited references

1. Tettelin, H., Masignani, V., Cieslewicz, M.J., Donati, C., Medini, D., Ward, N.L., Angiuoli, S.V., Crabtree, J., Jones, A.L., Durkin, A.S., et al. (2005). Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome". Proc Natl Acad Sci U S A *102*, 13950-13955.
2. Perna, N.T., Plunkett, G., 3rd, Burland, V., Mau, B., Glasner, J.D., Rose, D.J., Mayhew, G.F., Evans, P.S., Gregor, J., Kirkpatrick, H.A., et al. (2001). Genome sequence of enterohaemorrhagic Escherichia coli O157:H7. Nature *409*, 529-533.
3. Welch, R.A., Burland, V., Plunkett, G., 3rd, Redford, P., Roesch, P., Rasko, D., Buckles, E.L., Liou, S.R., Boutin, A., Hackett, J., et al. (2002). Extensive mosaic structure revealed by the complete genome sequence of uropathogenic Escherichia coli. Proc Natl Acad Sci U S A *99*, 17020-17024.
4. Dobrindt, U., Hochhut, B., Hentschel, U., and Hacker, J. (2004). Genomic islands in pathogenic and environmental microorganisms. Nat Rev Microbiol *2*, 414-424.
5. McInerney, J.O., McNally, A., and O'Connell, M.J. (2017). Why prokaryotes have pangenomes. Nat Microbiol *2*, 17040.
6. Martin, W.F. (2017). Too Much Eukaryote LGT. Bioessays *39*.
7. Leger, M.M., Eme, L., Stairs, C.W., and Roger, A.J. (2018). Demystifying Eukaryote Lateral Gene Transfer (Response to Martin 2017 DOI: 10.1002/bies.201700115). Bioessays *40*, e1700242.
8. Vos, M., Hesselman, M.C., Te Beek, T.A., van Passel, M.W.J., and Eyre-Walker, A. (2015). Rates of Lateral Gene Transfer in Prokaryotes: High but Why? Trends in microbiology *23*, 598-605.
9. Dubey, G.P., and Ben-Yehuda, S. (2011). Intercellular nanotubes mediate bacterial communication. Cell *144*, 590-600.
10. Fulsundar, S., Harms, K., Flaten, G.E., Johnsen, P.J., Chopade, B.A., and Nielsen, K.M. (2014). Gene transfer potential of outer membrane vesicles of Acinetobacter baylyi and effects of stress on vesiculation. Appl Environ Microbiol *80*, 3469-3483.
11. Beaber, J.W., and Waldor, M.K. (2004). Identification of operators and promoters that control SXT conjugative transfer. J Bacteriol *186*, 5945-5949.
12. Guerin, E., Cambray, G., Sanchez-Alberola, N., Campoy, S., Erill, I., Da Re, S., Gonzalez-Zorn, B., Barbe, J., Ploy, M.C., and Mazel, D. (2009). The SOS response controls integron recombination. Science *324*, 1034.
13. Nanda, A.M., Thormann, K., and Frunzke, J. (2015). Impact of spontaneous prophage induction on the fitness of bacterial populations and host-microbe interactions. J Bacteriol *197*, 410-419.
14. Twiss, E., Coros, A.M., Tavakoli, N.P., and Derbyshire, K.M. (2005). Transposition is modulated by a diverse set of host factors in Escherichia coli and is stimulated by nutritional stress. Mol Microbiol *57*, 1593-1607.
15. Stecher, B., Denzler, R., Maier, L., Bernet, F., Sanders, M.J., Pickard, D.J., Barthel, M., Westendorf, A.M., Krogfelt, K.A., Walker, A.W., et al. (2012). Gut inflammation can boost horizontal gene transfer between pathogenic and commensal Enterobacteriaceae. Proc Natl Acad Sci U S A *109*, 1269-1274.
16. Blokesch, M. (2016). Natural competence for transformation. Current biology : CB *26*, R1126-R1130.
17. Johnston, C., Martin, B., Fichant, G., Polard, P., and Claverys, J.P. (2014). Bacterial transformation: distribution, shared mechanisms and divergent control. Nat Rev Microbiol *12*, 181-196.
18. Koraimann, G., and Wagner, M.A. (2014). Social behavior and decision making in bacterial conjugation. Front Cell Infect Microbiol *4*, 54.
19. Seitz, P., and Blokesch, M. (2013). DNA-uptake machinery of naturally competent Vibrio cholerae. Proc Natl Acad Sci U S A *110*, 17987-17992.

20. Smillie, C.S., Smith, M.B., Friedman, J., Cordero, O.X., David, L.A., and Alm, E.J. (2011). Ecology drives a global network of gene exchange connecting the human microbiome. Nature *480*, 241-244.

21. Babic, A., Berkmen, M.B., Lee, C.A., and Grossman, A.D. (2011). Efficient gene transfer in bacterial cell chains. MBio *2*.

22. Hooper, S.D., Mavromatis, K., and Kyrpides, N.C. (2009). Microbial co-habitation and lateral gene transfer: what transposases can tell us. Genome Biol *10*, R45.

23. Chaffron, S., Rehrauer, H., Pernthaler, J., and von Mering, C. (2010). A global network of coexisting microbes from environmental and whole-genome sequence data. Genome Res *20*, 947-959.

24. Kloesges, T., Popa, O., Martin, W., and Dagan, T. (2011). Networks of gene sharing among 329 proteobacterial genomes reveal differences in lateral gene transfer frequency at different phylogenetic depths. Mol Biol Evol *28*, 1057-1074.

25. Popa, O., and Dagan, T. (2011). Trends and barriers to lateral gene transfer in prokaryotes. Curr Opin Microbiol *14*, 615-623.

26. Bolotin, E., and Hershberg, R. (2015). Gene Loss Dominates As a Source of Genetic Variation within Clonal Pathogenic Bacterial Species. Genome Biol Evol *7*, 2173-2187.

27. McNally, A., Thomson, N.R., Reuter, S., and Wren, B.W. (2016). 'Add, stir and reduce': Yersinia spp. as model bacteria for pathogen evolution. Nat Rev Microbiol *14*, 177-190.

28. Reuter, S., Connor, T.R., Barquist, L., Walker, D., Feltwell, T., Harris, S.R., Fookes, M., Hall, M.E., Petty, N.K., Fuchs, T.M., et al. (2014). Parallel independent evolution of pathogenicity within the genus Yersinia. Proc Natl Acad Sci U S A *111*, 6768-6773.

29. Reuter, S., Corander, J., de Been, M., Harris, S., Cheng, L., Hall, M., Thomson, N.R., and McNally, A. (2015). Directional gene flow and ecological separation in Yersinia enterocolitica. Microb Genom *1*, e000030.

30. Majewski, J., and Cohan, F.M. (1999). DNA sequence similarity requirements for interspecific recombination in Bacillus. Genetics *153*, 1525-1533.

31. Lovett, S.T., Hurley, R.L., Sutera, V.A., Jr., Aubuchon, R.H., and Lebedeva, M.A. (2002). Crossing over between regions of limited homology in Escherichia coli. RecA-dependent and RecA-independent pathways. Genetics *160*, 851-859.

32. Baharoglu, Z., Bikard, D., and Mazel, D. (2010). Conjugative DNA transfer induces the bacterial SOS response and promotes antibiotic resistance development through integron activation. PLoS Genet *6*, e1001165.

33. Cohan, F.M. (2017). Transmission in the Origins of Bacterial Diversity, From Ecotypes to Phyla. Microbiol Spectr *5*.

34. Sorek, R., Zhu, Y., Creevey, C.J., Francino, M.P., Bork, P., and Rubin, E.M. (2007). Genome-wide experimental determination of barriers to horizontal gene transfer. Science *318*, 1449-1452.

35. Tuller, T., Girshovich, Y., Sella, Y., Kreimer, A., Freilich, S., Kupiec, M., Gophna, U., and Ruppin, E. (2011). Association between translation efficiency and horizontal gene transfer within microbial communities. Nucleic Acids Res *39*, 4743-4755.

36. Hao, W., and Golding, G.B. (2006). The fate of laterally transferred genes: life in the fast lane to adaptation or death. Genome Res *16*, 636-643.

37. Marri, P.R., Hao, W., and Golding, G.B. (2007). The role of laterally transferred genes in adaptive evolution. BMC evolutionary biology *7 Suppl 1*, S8.

38. Porse, A., Schou, T.S., Munck, C., Ellabaan, M.M.H., and Sommer, M.O.A. (2018). Biochemical mechanisms determine the functional compatibility of heterologous genes. Nat Commun *9*, 522.

39. Szabova, J., Ruzicka, P., Verner, Z., Hampl, V., and Lukes, J. (2011). Experimental examination of EFL and MATX eukaryotic horizontal gene transfers: coexistence of mutually exclusive transcripts predates functional rescue. Mol Biol Evol *28*, 2371-2378.

40. Pal, C., Papp, B., and Lercher, M.J. (2005). Adaptive evolution of bacterial metabolic networks by horizontal gene transfer. Nat Genet *37*, 1372-1375.

41. Rivera, M.C., Jain, R., Moore, J.E., and Lake, J.A. (1998). Genomic evidence for two functionally distinct gene classes. Proc Natl Acad Sci U S A *95*, 6239-6244.

42. Cohen, O., Gophna, U., and Pupko, T. (2011). The complexity hypothesis revisited: connectivity rather than function constitutes a barrier to horizontal gene transfer. Mol Biol Evol *28*, 1481-1489.

43. Jain, R., Rivera, M.C., and Lake, J.A. (1999). Horizontal gene transfer among genomes: the complexity hypothesis. Proc Natl Acad Sci U S A *96*, 3801-3806.

44. Baltrus, D.A. (2013). Exploring the costs of horizontal gene transfer. Trends in Ecology & Evolution *28*, 489-495.

45. San Millan, A., Toll-Riera, M., Qi, Q., and MacLean, R.C. (2015). Interactions between horizontally acquired genes create a fitness cost in Pseudomonas aeruginosa. Nat Commun *6*, 6845.

46. Harrison, E., and Brockhurst, M.A. (2012). Plasmid-mediated horizontal gene transfer is a coevolutionary process. Trends in Microbiology *20*, 262-267.

47. Stevenson, C., Hall, J.P., Harrison, E., Wood, A., and Brockhurst, M.A. (2017). Gene mobility promotes the spread of resistance in bacterial populations. Isme J *11*, 1930-1932.

48. Brown, C.J., Sen, D., Yano, H., Bauer, M.L., Rogers, L.M., Van der Auwera, G.A., and Top, E.M. (2013). Diverse broad-host-range plasmids from freshwater carry few accessory genes. Appl Environ Microbiol *79*, 7684-7695.

49. Lopatkin, A.J., Meredith, H.R., Srimani, J.K., Pfeiffer, C., Durrett, R., and You, L. (2017). Persistence and reversal of plasmid-mediated antibiotic resistance. Nat Commun *8*, 1689.

50. Hall, J.P., Wood, A.J., Harrison, E., and Brockhurst, M.A. (2016). Source-sink plasmid transfer dynamics maintain gene mobility in soil bacterial communities. Proc Natl Acad Sci U S A *113*, 8260-8265.

51. Fox, R.E., Zhong, X., Krone, S.M., and Top, E.M. (2008). Spatial structure and nutrients promote invasion of IncP-1 plasmids in bacterial populations. Isme J *2*, 1024-1039.

52. Bahl, M.I., Hansen, L.H., and Sorensen, S.J. (2007). Impact of conjugal transfer on the stability of IncP-1 plasmid pKJK5 in bacterial populations. FEMS Microbiol Lett *266*, 250-256.

53. Lopatkin, A.J., Huang, S., Smith, R.P., Srimani, J.K., Sysoeva, T.A., Bewick, S., Karig, D.K., and You, L. (2016). Antibiotics as a selective driver for conjugation dynamics. Nat Microbiol *1*, 16044.

54. Hall, J.P.J., Williams, D., Paterson, S., Harrison, E., and Brockhurst, M.A. (2017). Positive selection inhibits gene mobilisation and transfer in soil bacterial communities. Nat Ecol Evol *1*, 1348-1353.

55. Gao, N.L., Zhang, C., Zhang, Z., Hu, S., Lercher, M.J., Zhao, X.M., Bork, P., Liu, Z., and Chen, W.H. (2018). MVP: a microbe-phage interaction database. Nucleic Acids Res *46*, D700-D707.

56. Hyman, P., and Abedon, S.T. (2010). Bacteriophage host range and bacterial resistance. Adv Appl Microbiol *70*, 217-248.

57. Jain, A., and Srivastava, P. (2013). Broad host range plasmids. FEMS Microbiol Lett *348*, 87-96.

58. Halary, S., Leigh, J.W., Cheaib, B., Lopez, P., and Bapteste, E. (2010). Network analyses structure genetic diversity in independent genetic worlds. Proc Natl Acad Sci U S A *107*, 127-132.

59. Sheppard, A.E., Stoesser, N., Wilson, D.J., Sebra, R., Kasarskis, A., Anson, L.W., Giess, A., Pankhurst, L.J., Vaughan, A., Grim, C.J., et al. (2016). Nested Russian Doll-Like Genetic Mobility Drives Rapid Dissemination of the Carbapenem Resistance Gene blaKPC. Antimicrob Agents Chemother *60*, 3767-3778.

60. He, S., Chandler, M., Varani, A.M., Hickman, A.B., Dekker, J.P., and Dyda, F. (2016). Mechanisms of Evolution in High-Consequence Drug Resistance Plasmids. MBio *7*.

61. Greated, A., Lambertsen, L., Williams, P.A., and Thomas, C.M. (2002). Complete sequence of the IncP-9 TOL plasmid pWW0 from Pseudomonas putida. Environ Microbiol *4*, 856-871.

62. Pesesky, M.W., Tilley, R., and Beck, D.A.C. (2019). Mosaic plasmids are abundant and unevenly distributed across prokaryotic taxa. Plasmid *102*, 10-18.

63. Nakamura, Y., Itoh, T., Matsuda, H., and Gojobori, T. (2004). Biased biological functions of horizontally transferred genes in prokaryotic genomes. Nat Genet *36*, 760-766.

64. Porse, A., Schonning, K., Munck, C., and Sommer, M.O.A. (2016). Survival and Evolution of a Large Multidrug Resistance Plasmid in New Clinical Bacterial Hosts. Mol Biol Evol *33*, 2860-2873.

65. Silva, J.B., Storms, Z., and Sauvageau, D. (2016). Host receptors for bacteriophage adsorption. Fems Microbiol Lett *363*.

66. San Millan, A., Pena-Miller, R., Toll-Riera, M., Halbert, Z.V., McLean, A.R., Cooper, B.S., and MacLean, R.C. (2014). Positive selection and compensatory adaptation interact to stabilize non-transmissible plasmids. Nat Commun *5*, 5208.

67. Harrison, E., Dytham, C., Hall, J.P., Guymer, D., Spiers, A.J., Paterson, S., and Brockhurst, M.A. (2016). Rapid compensatory evolution promotes the survival of conjugative plasmids. Mob Genet Elements *6*, e1179074.

68. Loftie-Eaton, W., Bashford, K., Quinn, H., Dong, K., Millstein, J., Hunter, S., Thomason, M.K., Merrikh, H., Ponciano, J.M., and Top, E.M. (2017). Compensatory mutations improve general permissiveness to antibiotic resistance plasmids. Nat Ecol Evol *1*, 1354-1363.

69. Harrison, E., Guymer, D., Spiers, A.J., Paterson, S., and Brockhurst, M.A. (2015). Parallel compensatory evolution stabilizes plasmids across the parasitism-mutualism continuum. Current biology : CB *25*, 2034-2039.

70. Yano, H., Wegrzyn, K., Loftie-Eaton, W., Johnson, J., Deckert, G.E., Rogers, L.M., Konieczny, I., and Top, E.M. (2016). Evolved plasmid-host interactions reduce plasmid interference cost. Mol Microbiol *101*, 743-756.

71. Bottery, M.J., Wood, A.J., and Brockhurst, M.A. (2019). Temporal dynamics of bacteria-plasmid coevolution under antibiotic selection. Isme J *13*, 559-562.

72. Bottery, M.J., Wood, A.J., and Brockhurst, M.A. (2017). Adaptive modulation of antibiotic resistance through intragenomic coevolution. Nat Ecol Evol *1*, 1364-1369.

73. Porse, A., Schonning, K., Munck, C., and Sommer, M.O. (2016). Survival and Evolution of a Large Multidrug Resistance Plasmid in New Clinical Bacterial Hosts. Mol Biol Evol *33*, 2860-2873.

74. Turner, P.E., Williams, E.S., Okeke, C., Cooper, V.S., Duffy, S., and Wertz, J.E. (2014). Antibiotic resistance correlates with transmission in plasmid evolution. Evolution *68*, 3368-3380.

75. Bobay, L.M., Touchon, M., and Rocha, E.P.C. (2014). Pervasive domestication of defective prophages by bacteria. P Natl Acad Sci USA *111*, 12127-12132.

76. Kottara, A., Hall, J.P.J., Harrison, E., and Brockhurst, M.A. (2018). Variable plasmid fitness effects and mobile genetic element dynamics across Pseudomonas species. FEMS microbiology ecology *94*.

77. Agrawal, N., Dasaradhi, P.V., Mohmmed, A., Malhotra, P., Bhatnagar, R.K., and Mukherjee, S.K. (2003). RNA interference: biology, mechanism, and applications. Microbiol Mol Biol Rev *67*, 657-685.

78. Lucchini, S., Rowley, G., Goldberg, M.D., Hurd, D., Harrison, M., and Hinton, J.C. (2006). H-NS mediates the silencing of laterally acquired genes in bacteria. Plos Pathog *2*, e81.

79. Marraffini, L.A., and Sontheimer, E.J. (2008). CRISPR interference limits horizontal gene transfer in staphylococci by targeting DNA. Science *322*, 1843-1845.

80. Dupuis, M.E., Villion, M., Magadan, A.H., and Moineau, S. (2013). CRISPR-Cas and restriction-modification systems are compatible and increase phage resistance. Nat Commun *4*, 2087.

81. Oliveira, P.H., Touchon, M., and Rocha, E.P. (2016). Regulation of genetic flux between bacteria by restriction-modification systems. Proc Natl Acad Sci U S A *113*, 5658-5663.

82. Palmer, K.L., and Gilmore, M.S. (2010). Multidrug-resistant enterococci lack CRISPR-cas. MBio *1*.

83. Watson, B.N.J., Staals, R.H.J., and Fineran, P.C. (2018). CRISPR-Cas-Mediated Phage Resistance Enhances Horizontal Gene Transfer by Transduction. MBio *9*.

84. Goldberg, G.W., Jiang, W., Bikard, D., and Marraffini, L.A. (2014). Conditional tolerance of temperate phages via transcription-dependent CRISPR-Cas targeting. Nature *514*, 633-637.

85. Faure, G., Makarova, K.S., and Koonin, E.V. (2019). CRISPR-Cas: Complex Functional Networks and Multiple Roles beyond Adaptive Immunity. J Mol Biol *431*, 3-20.

86. Gophna, U., Kristensen, D.M., Wolf, Y.I., Popa, O., Drevet, C., and Koonin, E.V. (2015). No evidence of inhibition of horizontal gene transfer by CRISPR-Cas on evolutionary timescales. Isme J *9*, 2021-2027.

87. Gao, N.L., Chen, J., Lercher, M.J., and Chen, W.-H. (2018). Prokaryotic genome expansion is facilitated by phages and plasmids but impaired by CRISPR. BioRxiv.

88. Doron, S., Melamed, S., Ofir, G., Leavitt, A., Lopatina, A., Keren, M., Amitai, G., and Sorek, R. (2018). Systematic discovery of antiphage defense systems in the microbial pangenome. Science *359*.

89. Thomas, C.M., and Nielsen, K.M. (2005). Mechanisms of, and barriers to, horizontal gene transfer between bacteria. Nat Rev Microbiol *3*, 711-721.

90. Berngruber, T.W., Weissing, F.J., and Gandon, S. (2010). Inhibition of superinfection and the evolution of viral latency. J Virol *84*, 10200-10208.

91. Faure, G., Shmakov, S.A., Yan, W.X., Cheng, D.R., Scott, D.A., Peters, J.E., Makarova, K.S., and Koonin, E.V. (2019). CRISPR-Cas in mobile genetic elements: counter-defence and beyond. Nat Rev Microbiol.

92. Hartl, D.L., and Clark, A.G. (2007). Principles of population genetics, 4th Edition, (Sunderland, Mass.: Sinauer Associates).

93. Charlesworth, B. (2009). Effective population size and patterns of molecular evolution and variation. Nat Rev Genet *10*, 195-205.

94. San Millan, A., and MacLean, R.C. (2017). Fitness Costs of Plasmids: a Limit to Plasmid Transmission. Microbiol Spectr *5*.

95. Vogwill, T., and MacLean, R.C. (2015). The genetic basis of the fitness costs of antimicrobial resistance: a meta-analysis approach. Evol Appl *8*, 284-295.

96. Price, M.N., Wetmore, K.M., Waters, R.J., Callaghan, M., Ray, J., Liu, H., Kuehl, J.V., Melnyk, R.A., Lamson, J.S., Suh, Y., et al. (2018). Mutant phenotypes for thousands of bacterial genes of unknown function. Nature *557*, 503-509.

97. van Opijnen, T., and Camilli, A. (2013). Transposon insertion sequencing: a new tool for systems-level analysis of microorganisms. Nat Rev Microbiol *11*, 435-442.

98. Giaever, G., Chu, A.M., Ni, L., Connelly, C., Riles, L., Veronneau, S., Dow, S., Lucau-Danila, A., Anderson, K., Andre, B., et al. (2002). Functional profiling of the Saccharomyces cerevisiae genome. Nature *418*, 387-391.

99. Sela, I., Wolf, Y.I., and Koonin, E.V. (2016). Theory of prokaryotic genome evolution. P Natl Acad Sci USA *113*, 11399-11407.

100. Bobay, L.M., and Ochman, H. (2018). Factors driving effective population size and pan-genome evolution in bacteria. Bmc Evolutionary Biology *18*.

101. Mira, A., Ochman, H., and Moran, N.A. (2001). Deletional bias and the evolution of bacterial genomes. Trends Genet *17*, 589-596.

102. Gil, R., Sabater-Munoz, B., Latorre, A., Silva, F.J., and Moya, A. (2002). Extreme genome reduction in Buchnera spp.: toward the minimal genome needed for symbiotic life. Proc Natl Acad Sci U S A *99*, 4454-4458.

103. Veyrier, F.J., Dufort, A., and Behr, M.A. (2011). The rise and fall of the Mycobacterium tuberculosis genome. Trends Microbiol *19*, 156-161.

104. Niehus, R., Mitri, S., Fletcher, A.G., and Foster, K.R. (2015). Migration and horizontal gene transfer divide microbial genomes into multiple niches. Nature Communications *6*.

105. Fraser, C., Alm, E.J., Polz, M.F., Spratt, B.G., and Hanage, W.P. (2009). The bacterial species challenge: making sense of genetic and ecological diversity. Science *323*, 741-746.

106. Andreani, N.A., Hesse, E., and Vos, M. (2017). Prokaryote genome fluidity is dependent on effective population size. Isme J *11*, 1719-1721.

107. 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.

108. Baumdicker, F., Hess, W.R., and Pfaffelhuber, P. (2012). The Infinitely Many Genes Model for the Distributed Genome of Bacteria. Genome Biol Evol *4*, 443-456.

109. Rocha, E.P.C. (2018). Neutral Theory, Microbial Practice: Challenges in Bacterial Population Genetics. Mol Biol Evol *35*, 1338-1347.

110. May, R.M., and Anderson, R.M. (1983). Epidemiology and Genetics in the Coevolution of Parasites and Hosts. P Roy Soc Lond a Mat *390*, 219-219.

111. Partridge, S.R., Kwong, S.M., Firth, N., and Jensen, S.O. (2018). Mobile Genetic Elements Associated with Antimicrobial Resistance. Clin Microbiol Rev *31*.

112. Rozwandowicz, M., Brouwer, M.S.M., Fischer, J., Wagenaar, J.A., Gonzalez-Zorn, B., Guerra, B., Mevius, D.J., and Hordijk, J. (2018). Plasmids carrying antimicrobial resistance genes in Enterobacteriaceae. J Antimicrob Chemother *73*, 1121-1137.

113. Cordero, O.X., and Polz, M.F. (2014). Explaining microbial genomic diversity in light of evolutionary ecology. Nat Rev Microbiol *12*, 263-273.

114. Gardner, A. (2009). Adaptation as organism design. Biology letters *5*, 861-864.

115. Karkman, A., Parnanen, K., and Larsson, D.G.J. (2019). Fecal pollution can explain antibiotic resistance gene abundances in anthropogenically impacted environments. Nat Commun *10*, 80.

116. Klumper, U., Riber, L., Dechesne, A., Sannazzarro, A., Hansen, L.H., Sorensen, S.J., and Smets, B.F. (2015). Broad host range plasmids can invade an unexpectedly diverse fraction of a soil bacterial community. Isme J *9*, 934-945.

117. Datta, M.S., Sliwerska, E., Gore, J., Polz, M.F., and Cordero, O.X. (2016). Microbial interactions lead to rapid micro-scale successions on model marine particles. Nat Commun *7*, 11965.

118. Leibold, M.A. (2018). Metacommunity ecology, (Princeton, NJ: Princeton University Press).

119. Makarova, K.S., Wolf, Y.I., Mekhedov, S.L., Mirkin, B.G., and Koonin, E.V. (2005). Ancestral paralogs and pseudoparalogs and their role in the emergence of the eukaryotic cell. Nucleic Acids Res *33*, 4626-4638.

120. Doolittle, R.F. (1995). The multiplicity of domains in proteins. Annu Rev Biochem *64*, 287-314.

121. Nevoigt, E., Fassbender, A., and Stahl, U. (2000). Cells of the yeast Saccharomyces cerevisiae are transformable by DNA under non-artificial conditions. Yeast *16*, 1107-1110.

122. Heinemann, J.A., and Sprague, G.F., Jr. (1989). Bacterial conjugative plasmids mobilize DNA transfer between bacteria and yeast. Nature *340*, 205-209.

123. Soanes, D., and Richards, T.A. (2014). Horizontal gene transfer in eukaryotic plant pathogens. Annu Rev Phytopathol *52*, 583-614.

124. Monier, A., Pagarete, A., de Vargas, C., Allen, M.J., Read, B., Claverie, J.M., and Ogata, H. (2009). Horizontal gene transfer of an entire metabolic pathway between a eukaryotic alga and its DNA virus. Genome Res *19*, 1441-1449.

125. Gallot-Lavallee, L., and Blanc, G. (2017). A Glimpse of Nucleo-Cytoplasmic Large DNA Virus Biodiversity through the Eukaryotic Genomics Window. Viruses *9*.

126. Kim, G., LeBlanc, M.L., Wafula, E.K., dePamphilis, C.W., and Westwood, J.H. (2014). Plant science. Genomic-scale exchange of mRNA between a parasitic plant and its hosts. Science *345*, 808-811.

127. Doolittle, W.F. (1998). You are what you eat: a gene transfer ratchet could account for bacterial genes in eukaryotic nuclear genomes. Trends Genet *14*, 307-311.

128. Glass, N.L., Jacobson, D.J., and Shiu, P.K. (2000). The genetics of hyphal fusion and vegetative incompatibility in filamentous ascomycete fungi. Annu Rev Genet *34*, 165-186.

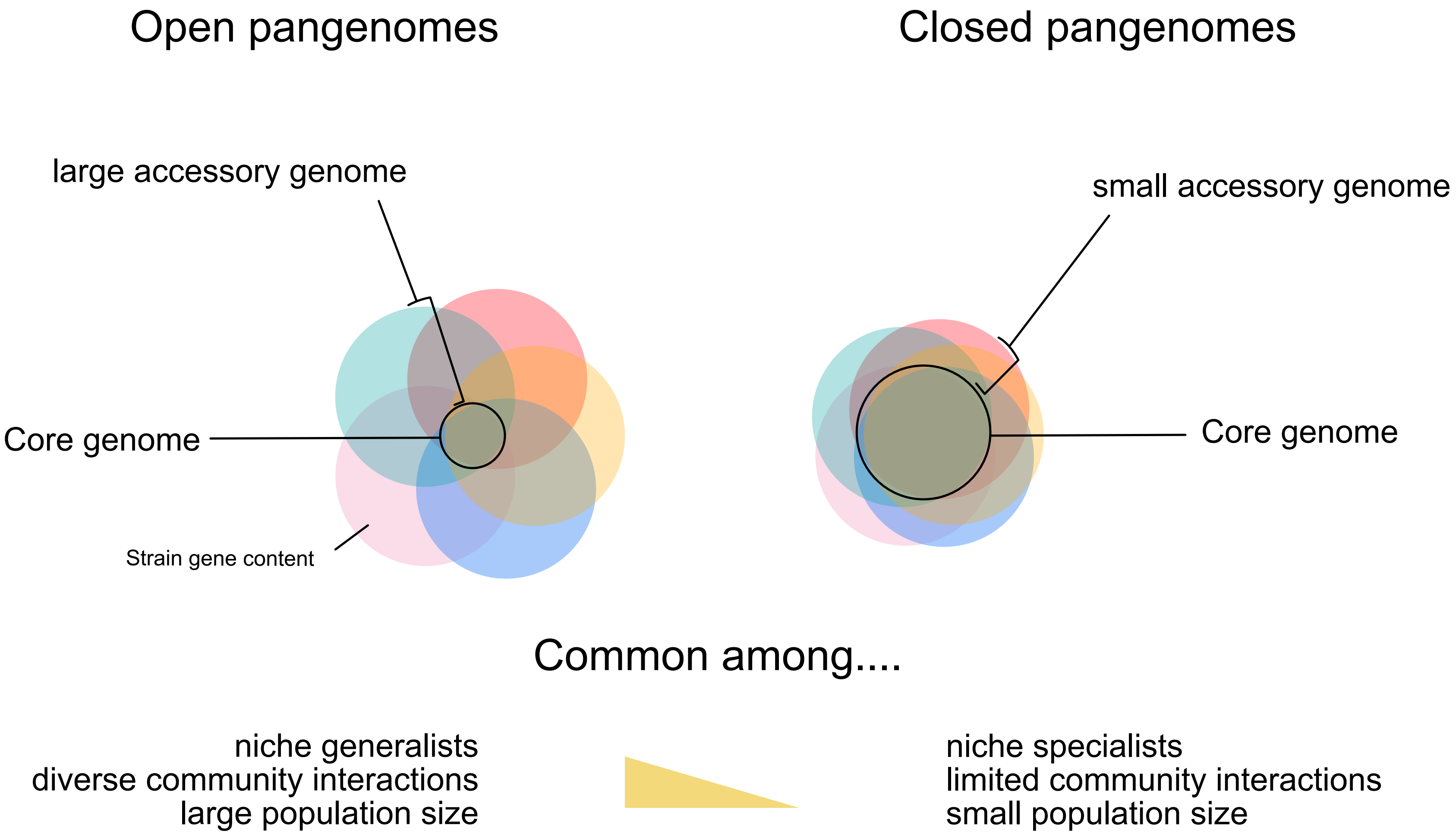
129. McCarthy, C.G.P., and Fitzpatrick, D.A. (2019). Pan-genome analyses of model fungal species. Microb Genom *5*.

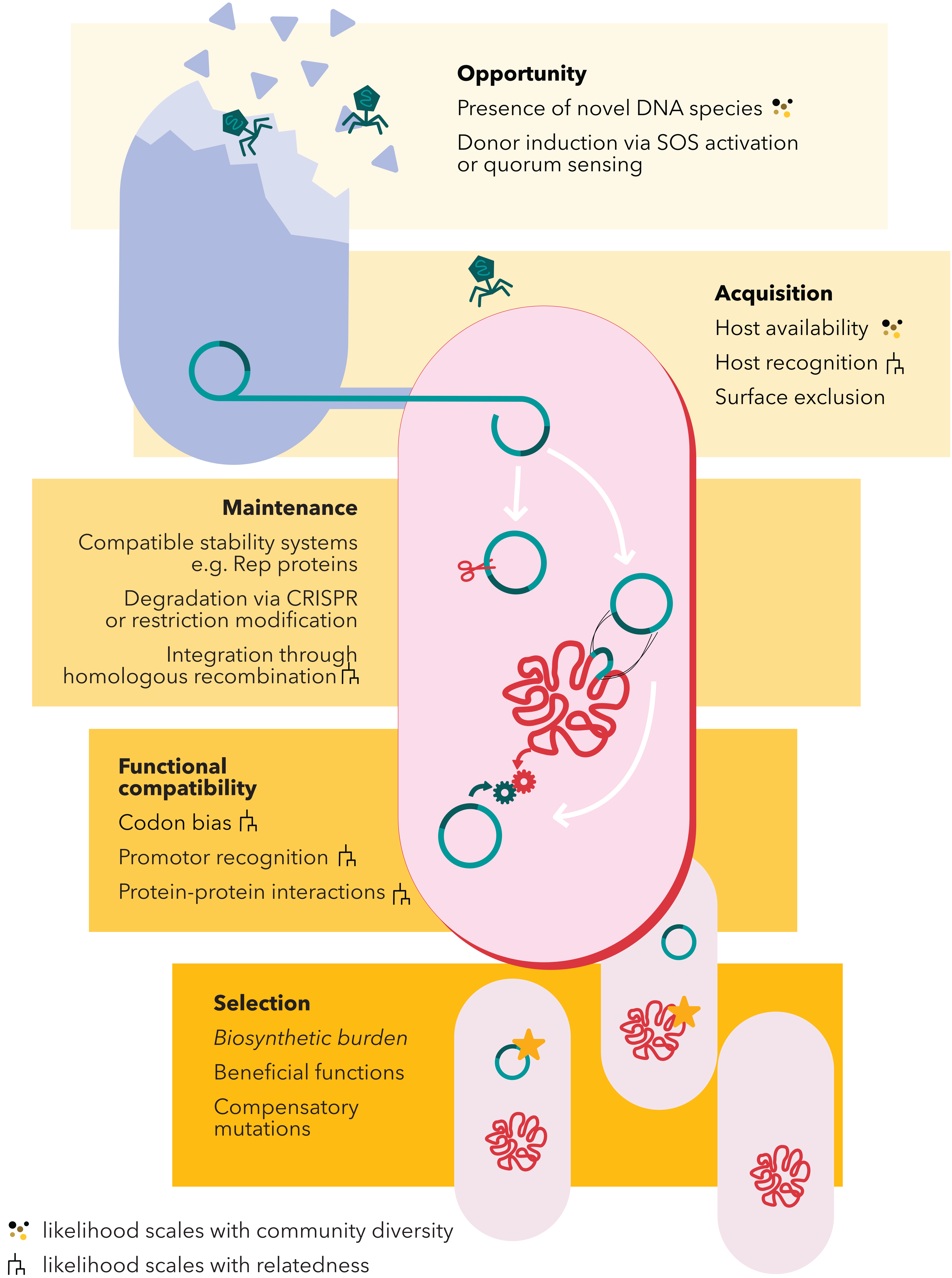
130. Read, B.A., Kegel, J., Klute, M.J., Kuo, A., Lefebvre, S.C., Maumus, F., Mayer, C., Miller, J., Monier, A., Salamov, A., et al. (2013). Pan genome of the phytoplankton Emiliania underpins its global distribution. Nature *499*, 209-213.

131. Temporini, E.D., and VanEtten, H.D. (2004). An analysis of the phylogenetic distribution of the pea pathogenicity genes of Nectria haematococca MPVI supports the hypothesis of their origin by horizontal transfer and uncovers a potentially new pathogen of garden pea: Neocosmospora boniensis. Curr Genet *46*, 29-36.

132. Coleman, J.J., Rounsley, S.D., Rodriguez-Carres, M., Kuo, A., Wasmann, C.C., Grimwood, J., Schmutz, J., Taga, M., White, G.J., Zhou, S., et al. (2009). The genome of Nectria haematococca: contribution of supernumerary chromosomes to gene expansion. PLoS Genet *5*, e1000618.

133. He, C., Rusu, A.G., Poplawski, A.M., Irwin, J.A., and Manners, J.M. (1998). Transfer of a supernumerary chromosome between vegetatively incompatible biotypes of the fungus Colletotrichum gloeosporioides. Genetics *150*, 1459-1466.


Open pangenomes
Closed pangenomes
large accessory genome
Core genome
Strain gene content
small accessory genome
Core genome
Common among....
niche generalists
diverse community interactions
large population size
niche specialists
limited community interactions
small population size



Opportunity

Presence of novel DNA species ⁘

Donor induction via SOS activation or quorum sensing

Acquisition

Host availability ⁘

Host recognition ⑂

Surface exclusion

Maintenance

Compatible stability systems e.g. Rep proteins

Degradation via CRISPR or restriction modification

Integration through homologous recombination ⑂

Functional compatibility

Codon bias ⑂

Promotor recognition ⑂

Protein-protein interactions ⑂

Selection

Biosynthetic burden

Beneficial functions

Compensatory mutations

⁘ likelihood scales with community diversity

⑂ likelihood scales with relatedness