# 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

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

54 The current challenge is to move beyond this phenomenological description of pangenomes
55 to forge an understanding of the mechanisms and processes that determine their structure. A
56 genome sequence is a snapshot of a strain in time. Some of the genes and mutations in that
57 snapshot share a long history and are destined to remain associated, while other members
58 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*

86 The proximal environmental triggers activating expression of gene exchange machinery vary
87 between systems and with different species, but some common themes can be identified. One
88 of these is stress. For example, the SOS response to DNA damage, triggered by some
89 antibiotics, reactive oxygen, and UV radiation, activates transfer of the Vibrio cholerae STX
90 element [11], causes integron rearrangement [12], and activates integrated bacteriophage
91 [13]. Transposons in *E. coli* become active under nutritional stress [14], plasmid conjugation
92 rates are increased in response to host inflammation in mammalian gut [15], and starvation
93 conditions activate natural competence [16]. However, different stress responses can have
94 divergent effects in different species [17], and donors, recipients, and mobile genetic elements
95 may each have their own cues. For example, some mobile genetic elements, such as the
96 pheromone-inducible conjugative plasmids of *Enterococcus*, have evolved mechanisms to
97 detect the presence of recipients [18], and transformation is induced by quorum sensing and
98 by specific nutrients in some species of *Vibrio* [19].

99 Ecology appears to be a principal determinant of gene-sharing [20], suggesting that the
100 transfer of genes is to some extent limited by ecological opportunity and occupancy of shared
101 habitats. Several gene transfer mechanisms including conjugation and nanotubes require
102 close physical proximity and thus HGT is probabilistically likely to be most efficient between
103 immediate neighbours [21]. Consequently, the size of the gene pool from which a species can
104 draw will be dependent on the diversity of environments they occupy as well as the community
105 diversity these contain. Correspondingly, networks of gene sharing have shown that co-
106 occurrence of species in a habitat increases the probability of gene sharing [22-25]. Niche
107 specialists likely to exist in stable environments with very low diversity, such as endosymbionts
108 [24], have more closed pan-genomes than those that exist in diverse communities and more
109 variable environments.

110 Among symbionts and pathogens with low rates of gene gain through HGT, variation in gene
111 loss among lineages can be the primary cause of diversity among clonal lineages, and can
112 lead to large phenotypic differences [26]. Whereas gene loss can be positively selected in

113 large populations with efficient selection, in intracellular symbionts and pathogens with low
114 effective population size, gene loss is more likely to be a result of relaxed selection and drift
115 [27]. How the balance of gene gain and loss contributes to the formation of a pangenome is
116 well-illustrated by *Yersinia enterocolitica*. The species is composed of five phylogenetically
117 distinct groups, four of which are pathogenic to humans and have emerged from a non-
118 pathogenic ancestor, driven by a single acquisition of a large virulence plasmid [28]. Following
119 plasmid acquisition, the splits between the four pathogenic groups are delineated at a
120 pangenome level by differential losses of genes present in the ancestor, alongside HGTs
121 leading to switches in serotype [29].

122

123 *Mechanistic drivers and barriers of HGT*

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

133 Genes must also be transferable and able to function in the host in order to have a phenotypic
134 effect visible to selection [33], which is dependent on recognition of promoters allowing for
135 gene expression [34], and comparable GC content, codon usage and compatible genetic
136 codes allowing for efficient translation [35], and in the case of DNA transfer between eukaryotic
137 genomes effective splicing of introns. Newly acquired genes evolve faster than older genes in
138 the same genome, potentially because of adaptation to their new genomic context [36, 37]. As
139 a general principle, many of these processes become more challenging across larger genetic

140 distances [38]. Correspondingly gene sharing has been shown to be most common between
141 closer phylogenetic relatives [25], which enhances both the likelihood of the transfer event and
142 the compatibility of genes between donor and recipient.

143 Mechanistic limitations are also likely to define the types of genes that are more readily shared,
144 and therefore more likely to contribute to the accessory genome. Incoming DNA can disrupt
145 cellular processes leading to severe fitness costs, and these genes are likely to be rapidly lost
146 from the population by purifying selection. Genes encoding core cellular functions, such as
147 those associated with transcription and translation, can be highly toxic when expressed in
148 foreign hosts [34, 39] and are poorly represented among horizontally transferred genes [40,
149 41]. This strong incompatibility may be due to disruption of or failure to maintain the large
150 number of protein-protein interactions that the protein must engage in to properly function.
151 Genes embedded within more complex interaction networks are therefore more disruptive and
152 less likely to maintain the necessary functional interaction network when transferred, a
153 phenomenon termed the complexity hypothesis [42, 43]. Mobile genetic elements (MGEs)
154 themselves are often associated with significant fitness costs that are caused by a range of
155 factors, including the biosynthetic cost of maintaining and expressing additional DNA, toxic
156 gene products, and epistasis between chromosomal and MGE-encoded genes [44]. This
157 disruptive effect of HGT is not surprising from an evolutionary perspective: HGT brings
158 together genes that have different evolutionary histories, and there is no a priori reason to
159 expect that these genes should function together harmoniously [45].

160

161 *Evolutionary conflict and collaboration in the pangenome*

162 Many of the mechanisms for horizontal gene transfer are encoded by infectious MGEs such
163 as viruses, plasmids, and transposable elements. Therefore, pangenomes are composites of
164 the host chromosome(s) together with MGEs that may be shared with other species. MGEs
165 encode accessory genes that may represent adaptive additions to the pangenome (e.g. by
166 providing a new ecological function or access to an otherwise inaccessible niche), but also

167 encode genes for selfish MGE-directed functions such as replication and transmission, as well
168 as many genes of unknown function. As semi-autonomous evolving entities we should expect
169 MGEs to maximise their own fitness through both vertical and horizontal transmission [46].
170 Encoding beneficial accessory genes can increase MGE fitness through enhanced vertical
171 transmission as positive selection drives clonal expansion [47]. However, being beneficial is
172 not necessary for MGE success. Many environmental plasmids do not encode any obvious
173 accessory genes [48] and are therefore likely to be genetic parasites. Experimental studies
174 show that high rates of horizontal transmission through conjugation can maintain costly
175 resistance plasmids in the absence of positive selection [47, 49, 50], and non-beneficial
176 plasmids can invade biofilm populations [51, 52]. Indeed, experiments with antibiotic
177 resistance and mercury detoxification plasmids have shown that positive selection for these
178 functions can limit their horizontal transfer by reducing the availability of recipient cells [47,
179 53]. Although, in the long run, purely infectious elements would be expected to become
180 increasingly efficient parasites by shedding their accessory genes, mobile genetic elements
181 that persist through horizontal transmission are likely to be especially prone to mediating gene
182 exchange [54]. Higher rates of horizontal transmission expose these MGEs to a wider diversity
183 of genomic environments, offering greater opportunity for other MGEs (e.g., transposons) to
184 integrate and hitch a ride. This inherent nestedness of pangenomes means that potentially
185 conflicting selective pressures may operate at different levels of complexity (e.g., at the level
186 of the gene, MGE, genome, population, and species etc.).

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

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

206 Because they can replicate by both vertical and horizontal transmission, MGEs can have
207 fitness interests that do not necessarily align with those of other parts of the (vertically-
208 inherited) genome. These 'divided loyalties' manifest in the fitness costs associated with MGE
209 acquisition and horizontal transmission, and result in intragenomic conflict. For example, while
210 conjugation provides an efficient mechanism for plasmids to transfer between bacteria, the
211 expression of conjugative machinery imposes a biosynthetic fitness cost on the donor cell [64],
212 and leaves the donor cell open to predation by pilus-targeting phage [65]. Resolution of host-
213 MGE conflict frequently requires compensatory mutation(s) to the MGE or the chromosome to
214 reduce the fitness costs of the newly acquired genes [46], which is promoted by positive
215 selection for MGE-encoded functions since this increases the population size and mutation
216 supply for MGE-carriers [66, 67]. Diverse compensatory mechanisms have been identified to
217 stabilise plasmids, but two common routes are mutations affecting host gene regulatory
218 networks [68, 69] or plasmid replication [45, 70]. By stabilising MGEs within bacterial lineages,
219 compensatory evolution can set the stage for more extensive coevolution between the MGE
220 and chromosome, driving reciprocal adaptations and counter-adaptations [46]. For example,

221 bacteria-plasmid coevolution rapidly led to the emergence of co-dependence of chromosomal
222 and plasmid replicons under antibiotic selection, together providing high-level resistance but
223 separately providing inadequate levels of resistance to persist in the environment they evolved
224 in [71, 72]. Compensation and coevolution can, in turn, drive the complete domestication of
225 MGEs and their integration into a more exclusively vertical mode of replication. In practice,
226 domestication involves downregulation, inactivation, or loss of the machinery involved in
227 horizontal transmission [73, 74]. For example, bacterial genomes contain numerous
228 prophages, some of which are incapable of horizontal transmission and now serve their
229 bacterial hosts as anti-competitor toxins [75]. Alternatively, recombination can relocate mobile
230 genes to less-mobile parts of the genome, e.g. chromosomal capture of resistance genes from
231 plasmids, a process rapid enough to be readily observable in the laboratory [50, 69, 76]. In so
232 doing, the signatures of gene acquisition are gradually lost from the genome sequence,
233 potentially explaining why many accessory genes originally transferred by an MGE are no
234 longer obviously associated with MGEs.

235

236 *Resisting HGT*

237 Due to the potential for conflict between MGEs and the host chromosome, immunity systems
238 which actively target incoming foreign DNA are widespread across eukaryotes and
239 prokaryotes. Systems exist in both eukaryotes (e.g. RNAi [77]) and prokaryotes (e.g. H-NS
240 [78]) to silence gene expression from foreign DNA. In prokaryotes CRISPR-Cas systems and
241 restriction-modification (R-M) systems target novel DNA for degradation, and can be an
242 effective defence against MGEs, potentially reducing HGT [79, 80]. A comparative analysis of
243 79 prokaryote genomes show that R-M systems structure gene sharing by favouring
244 exchanges between genomes with similar R-M systems [81]. The relationship between HGT
245 and CRISPR-Cas systems appears more complex: There are well-described cases where
246 CRISPR-Cas systems are negatively associated with MGE carriage within a species [82], but
247 CRISPR-Cas can also promote HGT in some cases [83]. Type-III CRISPR-Cas systems target

248 actively transcribed DNA via spacers derived from RNA transcripts [84] and may therefore be
249 more effective against phages and plasmids than DNA acquired by transformation [85]. Over
250 broader taxonomic scales, however, the correlation between CRISPR-Cas systems and the
251 rate of HGT is less clear and deserves further study [86, 87]. It is likely that additional
252 mechanisms for resisting gene acquisition will continue to be discovered [88]. Resistance
253 mechanisms protecting cells against incoming DNA can also be encoded by MGEs
254 themselves, highlighting how conflict between MGE could act to limit HGT. Both plasmids and
255 phages defend their host cells against super-infection though self-exclusion mechanisms [89,
256 90] and can encode their own CRISPR-Cas systems with spacer sequences targeting other
257 MGEs [91].

258

259 **How and why do pangenomes evolve?**

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

269

270 *Population genetic approaches to analysing the pangenome*

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

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

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

300 and genetically distinct subpopulations (or ecotypes) become sufficiently diverged to be
301 considered multiple, different species each with their own pangenome is contentious [33, 105].

302 Other studies using population genetics have questioned the role of selection in shaping the
303 pangenome. Comparing levels of synonymous nucleotide diversity, a surrogate measure of
304 $N_e$, with a measure pangenome fluidity showed a positive correlation between $N_e$ and
305 pangenome fluidity, that could arise because genetic drift leads to the loss of effectively neutral
306 accessory genes in species with low $N_e$ [106]. Further support for this idea comes from
307 comparing the observed distribution of gene frequencies in the pangenome with an expected
308 distribution generated by a neutral model. This approach, inspired by the infinite alleles model,
309 assumes that bacteria gain genes from an infinite pool of horizontally transferred genes and
310 subsequently lose these genes through drift [107, 108]. Accessory genes show a distribution
311 that is close to the expectations of a neutral model for widely distributed marine bacteria, but
312 with deviations that are consistent with selection shaping the pangenome [108]. It is unclear,
313 however, that currently available genomic data provide the necessary breadth and depth of
314 ecological sampling to adequately test these models.

315

316 *The limits of a population genetic approach*

317 Population genetics theory provides some simple guiding principles for understanding the
318 pangenome, but there are also potential difficulties with applying these models to understand
319 the pangenome [109]. For example, classical population genetic tests for selection rely on
320 comparing observed patterns of genetic polymorphisms and divergence with expected
321 patterns from a neutral model where evolution is driven by mutation and drift, but not selection.
322 Neutral models in population genetics assume that mutations at different sites in the genome
323 are not linked. This is a justifiable assumption in eukaryotic species with obligate sexual
324 reproduction, but the pangenome changes through the gain and loss of blocks of genes, for
325 example because they are all encoded on a MGE. An important consequence of this is that
326 strong selection for one gene (e.g. an antibiotic resistance gene) can lead to the spread of

327 linked mildly deleterious genes by co-selection, if there is a net fitness benefit of the MGE.
328 Similarly, genes that are linked to addiction systems, such as toxin-antitoxin systems, can be
329 maintained in populations by the toxic effects of MGE loss. In a broader perspective, the strong
330 linkage disequilibrium observed in clonal bacterial species means that there might be no
331 effectively neutral variation [109].

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

349

350 **Perspective**

351 Short-read sequencing technologies have produced a rapid accumulation of sequence data,
352 revealing the ubiquity and extent of pangenomes, especially in prokaryotes. At present,
353 however, we lack a unified theory to understand the forces structuring pangenomes, and this

354 will probably require the development of new theory that links together concepts from
355 evolutionary ecology and population genetics. To achieve this, there are some important
356 obstacles that need to be overcome:

357 • Defining the concept of pangenome adaptation: Adaptation is the "process of optimisation
358 of the phenotype under the action of natural selection" [114]. As a pangenome emerges
359 as an analytical result from comparing multiple genomes, we must take care when
360 specifying what adaptation means in this context, i.e. who or what is being optimised. While
361 a pangenome *can* contain adaptive genes that are transferred between species, the
362 pangenome does not evolve *for the purposes of* maintaining a pool of niche-adaptive
363 genes. Instead, its contents are defined by selection occurring at lower organisational
364 levels: the individual bacterial lineage that has acquired locally-beneficial genes, and the
365 persistent MGE. Neither does a broadly adaptive pangenome imply that the accessory
366 genes in a given genome are beneficial to that strain. Recent migration or gene acquisition
367 can result in a strain carrying neutral or deleterious genes which have not yet been lost
368 [115]. Finally, if the pangenome is defined as the sum-total of all genes in a species,
369 improved sequencing resolution will increasingly capture transient events which are
370 unlikely to be adaptive, inflating the size of the pangenome but diluting the signal of
371 adaptation. Enhanced biological insight into the gene function, as well as bioinformatic
372 tools that help us distinguish between transient associations and longer-term partnerships,
373 will guard us from incorrectly inferring adaptation in such instances.

374 • Measuring the rates of HGT in nature: The rate of horizontal gene transfer is key to both
375 the population genetic and eco-evolutionary perspectives on the pangenome, but our
376 knowledge of rate of HGT in the wild remains very limited. It might be possible to measure
377 these rate by using statistical methods to infer rates of HGT from genomic data, and
378 experimental methods that allow the spread of genes to be measured under natural
379 communities in real time using for example microcosm experiments [54, 116].

380 • Sampling genomes at ecologically-relevant scales: Microbial genomes are being
381 sequenced at an incredible rate, but it is very challenging to understand sequence data in
382 a population genetics context, there are often huge sampling biases in microbial sequence
383 datasets (intensive sampling of clinical outbreaks is the most extreme example). Given the
384 vast population size of microbes, we will only ever be able to achieve very sparse sampling
385 of microbial genomes, even with the most ambitious sequencing projects. We therefore
386 need to develop approaches to identify and sample ecologically coherent microbial
387 populations [113] or ecotypes [33]. For example, it is clear that some microbial populations
388 are structured at an incredibly fine scale, such as individual particles of detritus [117], and
389 this structuring can play a key role in the evolution of the pangenome [104]. Comparing a
390 small number of bacterial genomes sampled from many niches is likely to produce an
391 abundance of rare accessory genes, but these could either represent adaptive accessory
392 genes that are locally abundant but globally rare, or deleterious accessory genes that are
393 both locally and globally rare. One key technological development that may help with this
394 problem is to move from sequencing the genomes of bacterial isolates to single-cell
395 sequencing of bacteria from environmental samples.

396 • Developing eco-evolutionary models of pangenome evolution: The neutral theory of
397 molecular evolution has been so useful in revealing the action of natural selection because
398 it makes quantitative and falsifiable predictions that be tested by comparing datasets.
399 Given the complexity of forces shaping the pangenome it may be necessary to look outside
400 genetics for potential approaches: Pangenomes share many characteristics with
401 metacommunities, most notably the idea that entities (genes or species) are sampled from
402 a pool to form discrete sets (genomes or communities) that share biological cohesiveness
403 (pangenome or metacommunity). Metacommunity ecology has a well-developed body of
404 theory to understand how communities are assembled and structured [118], which may
405 help to unravel the processes causing the structure of pangenomes.

406

**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.

452 **Cited references**

453


454 1. Tettelin, H., Masignani, V., Cieslewicz, M.J., Donati, C., Medini, D., Ward, N.L.,
455 Angiuoli, S.V., Crabtree, J., Jones, A.L., Durkin, A.S., et al. (2005). Genome analysis
456 of multiple pathogenic isolates of Streptococcus agalactiae: implications for the
457 microbial "pan-genome". Proc Natl Acad Sci U S A *102*, 13950-13955.
458 2. Perna, N.T., Plunkett, G., 3rd, Burland, V., Mau, B., Glasner, J.D., Rose, D.J.,
459 Mayhew, G.F., Evans, P.S., Gregor, J., Kirkpatrick, H.A., et al. (2001). Genome
460 sequence of enterohaemorrhagic Escherichia coli O157:H7. Nature *409*, 529-533.
461 3. Welch, R.A., Burland, V., Plunkett, G., 3rd, Redford, P., Roesch, P., Rasko, D.,
462 Buckles, E.L., Liou, S.R., Boutin, A., Hackett, J., et al. (2002). Extensive mosaic
463 structure revealed by the complete genome sequence of uropathogenic Escherichia
464 coli. Proc Natl Acad Sci U S A *99*, 17020-17024.
465 4. Dobrindt, U., Hochhut, B., Hentschel, U., and Hacker, J. (2004). Genomic islands in
466 pathogenic and environmental microorganisms. Nat Rev Microbiol *2*, 414-424.
467 5. McInerney, J.O., McNally, A., and O'Connell, M.J. (2017). Why prokaryotes have
468 pangenomes. Nat Microbiol *2*, 17040.
469 6. Martin, W.F. (2017). Too Much Eukaryote LGT. Bioessays *39*.
470 7. Leger, M.M., Eme, L., Stairs, C.W., and Roger, A.J. (2018). Demystifying Eukaryote
471 Lateral Gene Transfer (Response to Martin 2017 DOI: 10.1002/bies.201700115).
472 Bioessays *40*, e1700242.
473 8. Vos, M., Hesselman, M.C., Te Beek, T.A., van Passel, M.W.J., and Eyre-Walker, A.
474 (2015). Rates of Lateral Gene Transfer in Prokaryotes: High but Why? Trends in
475 microbiology *23*, 598-605.
476 9. Dubey, G.P., and Ben-Yehuda, S. (2011). Intercellular nanotubes mediate bacterial
477 communication. Cell *144*, 590-600.
478 10. Fulsundar, S., Harms, K., Flaten, G.E., Johnsen, P.J., Chopade, B.A., and Nielsen,
479 K.M. (2014). Gene transfer potential of outer membrane vesicles of Acinetobacter
480 baylyi and effects of stress on vesiculation. Appl Environ Microbiol *80*, 3469-3483.
481 11. Beaber, J.W., and Waldor, M.K. (2004). Identification of operators and promoters that
482 control SXT conjugative transfer. J Bacteriol *186*, 5945-5949.
483 12. Guerin, E., Cambray, G., Sanchez-Alberola, N., Campoy, S., Erill, I., Da Re, S.,
484 Gonzalez-Zorn, B., Barbe, J., Ploy, M.C., and Mazel, D. (2009). The SOS response
485 controls integron recombination. Science *324*, 1034.
486 13. Nanda, A.M., Thormann, K., and Frunzke, J. (2015). Impact of spontaneous
487 prophage induction on the fitness of bacterial populations and host-microbe
488 interactions. J Bacteriol *197*, 410-419.
489 14. Twiss, E., Coros, A.M., Tavakoli, N.P., and Derbyshire, K.M. (2005). Transposition is
490 modulated by a diverse set of host factors in Escherichia coli and is stimulated by
491 nutritional stress. Mol Microbiol *57*, 1593-1607.
492 15. Stecher, B., Denzler, R., Maier, L., Bernet, F., Sanders, M.J., Pickard, D.J., Barthel,
493 M., Westendorf, A.M., Krogfelt, K.A., Walker, A.W., et al. (2012). Gut inflammation
494 can boost horizontal gene transfer between pathogenic and commensal
495 Enterobacteriaceae. Proc Natl Acad Sci U S A *109*, 1269-1274.
496 16. Blokesch, M. (2016). Natural competence for transformation. Current biology : CB *26*,
497 R1126-R1130.
498 17. Johnston, C., Martin, B., Fichant, G., Polard, P., and Claverys, J.P. (2014). Bacterial
499 transformation: distribution, shared mechanisms and divergent control. Nat Rev
500 Microbiol *12*, 181-196.
501 18. Koraimann, G., and Wagner, M.A. (2014). Social behavior and decision making in
502 bacterial conjugation. Front Cell Infect Microbiol *4*, 54.
503 19. Seitz, P., and Blokesch, M. (2013). DNA-uptake machinery of naturally competent
504 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.

559 40. Pal, C., Papp, B., and Lercher, M.J. (2005). Adaptive evolution of bacterial metabolic
560 networks by horizontal gene transfer. Nat Genet *37*, 1372-1375.
561 41. Rivera, M.C., Jain, R., Moore, J.E., and Lake, J.A. (1998). Genomic evidence for two
562 functionally distinct gene classes. Proc Natl Acad Sci U S A *95*, 6239-6244.
563 42. Cohen, O., Gophna, U., and Pupko, T. (2011). The complexity hypothesis revisited:
564 connectivity rather than function constitutes a barrier to horizontal gene transfer. Mol
565 Biol Evol *28*, 1481-1489.
566 43. Jain, R., Rivera, M.C., and Lake, J.A. (1999). Horizontal gene transfer among
567 genomes: the complexity hypothesis. Proc Natl Acad Sci U S A *96*, 3801-3806.
568 44. Baltrus, D.A. (2013). Exploring the costs of horizontal gene transfer. Trends in
569 Ecology & Evolution *28*, 489-495.
570 45. San Millan, A., Toll-Riera, M., Qi, Q., and MacLean, R.C. (2015). Interactions
571 between horizontally acquired genes create a fitness cost in Pseudomonas
572 aeruginosa. Nat Commun *6*, 6845.
573 46. Harrison, E., and Brockhurst, M.A. (2012). Plasmid-mediated horizontal gene transfer
574 is a coevolutionary process. Trends in Microbiology *20*, 262-267.
575 47. Stevenson, C., Hall, J.P., Harrison, E., Wood, A., and Brockhurst, M.A. (2017). Gene
576 mobility promotes the spread of resistance in bacterial populations. Isme J *11*, 1930-
577 1932.
578 48. Brown, C.J., Sen, D., Yano, H., Bauer, M.L., Rogers, L.M., Van der Auwera, G.A.,
579 and Top, E.M. (2013). Diverse broad-host-range plasmids from freshwater carry few
580 accessory genes. Appl Environ Microbiol *79*, 7684-7695.
581 49. Lopatkin, A.J., Meredith, H.R., Srimani, J.K., Pfeiffer, C., Durrett, R., and You, L.
582 (2017). Persistence and reversal of plasmid-mediated antibiotic resistance. Nat
583 Commun *8*, 1689.
584 50. Hall, J.P., Wood, A.J., Harrison, E., and Brockhurst, M.A. (2016). Source-sink
585 plasmid transfer dynamics maintain gene mobility in soil bacterial communities. Proc
586 Natl Acad Sci U S A *113*, 8260-8265.
587 51. Fox, R.E., Zhong, X., Krone, S.M., and Top, E.M. (2008). Spatial structure and
588 nutrients promote invasion of IncP-1 plasmids in bacterial populations. Isme J *2*,
589 1024-1039.
590 52. Bahl, M.I., Hansen, L.H., and Sorensen, S.J. (2007). Impact of conjugal transfer on
591 the stability of IncP-1 plasmid pKJK5 in bacterial populations. FEMS Microbiol Lett
592 *266*, 250-256.
593 53. Lopatkin, A.J., Huang, S., Smith, R.P., Srimani, J.K., Sysoeva, T.A., Bewick, S.,
594 Karig, D.K., and You, L. (2016). Antibiotics as a selective driver for conjugation
595 dynamics. Nat Microbiol *1*, 16044.
596 54. Hall, J.P.J., Williams, D., Paterson, S., Harrison, E., and Brockhurst, M.A. (2017).
597 Positive selection inhibits gene mobilisation and transfer in soil bacterial
598 communities. Nat Ecol Evol *1*, 1348-1353.
599 55. Gao, N.L., Zhang, C., Zhang, Z., Hu, S., Lercher, M.J., Zhao, X.M., Bork, P., Liu, Z.,
600 and Chen, W.H. (2018). MVP: a microbe-phage interaction database. Nucleic Acids
601 Res *46*, D700-D707.
602 56. Hyman, P., and Abedon, S.T. (2010). Bacteriophage host range and bacterial
603 resistance. Adv Appl Microbiol *70*, 217-248.
604 57. Jain, A., and Srivastava, P. (2013). Broad host range plasmids. FEMS Microbiol Lett
605 *348*, 87-96.
606 58. Halary, S., Leigh, J.W., Cheaib, B., Lopez, P., and Bapteste, E. (2010). Network
607 analyses structure genetic diversity in independent genetic worlds. Proc Natl Acad
608 Sci U S A *107*, 127-132.
609 59. Sheppard, A.E., Stoesser, N., Wilson, D.J., Sebra, R., Kasarskis, A., Anson, L.W.,
610 Giess, A., Pankhurst, L.J., Vaughan, A., Grim, C.J., et al. (2016). Nested Russian
611 Doll-Like Genetic Mobility Drives Rapid Dissemination of the Carbapenem
612 Resistance Gene blaKPC. Antimicrob Agents Chemother *60*, 3767-3778.

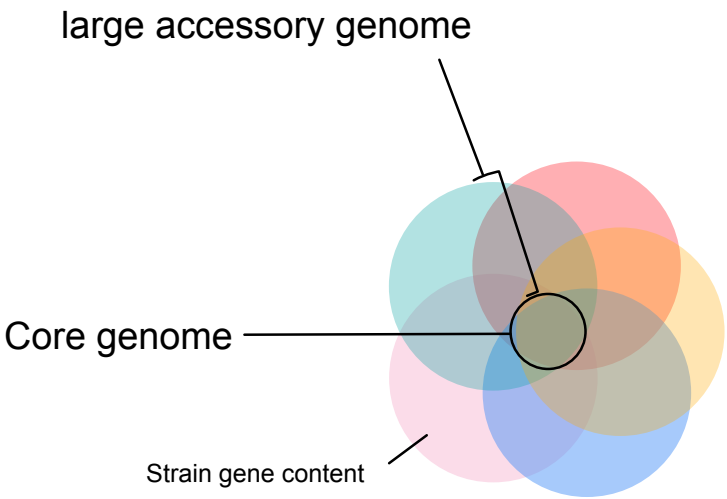
613 60. He, S., Chandler, M., Varani, A.M., Hickman, A.B., Dekker, J.P., and Dyda, F. (2016).
614 Mechanisms of Evolution in High-Consequence Drug Resistance Plasmids. MBio *7*.
615 61. Greated, A., Lambertsen, L., Williams, P.A., and Thomas, C.M. (2002). Complete
616 sequence of the IncP-9 TOL plasmid pWW0 from Pseudomonas putida. Environ
617 Microbiol *4*, 856-871.
618 62. Pesesky, M.W., Tilley, R., and Beck, D.A.C. (2019). Mosaic plasmids are abundant
619 and unevenly distributed across prokaryotic taxa. Plasmid *102*, 10-18.
620 63. Nakamura, Y., Itoh, T., Matsuda, H., and Gojobori, T. (2004). Biased biological
621 functions of horizontally transferred genes in prokaryotic genomes. Nat Genet *36*,
622 760-766.
623 64. Porse, A., Schonning, K., Munck, C., and Sommer, M.O.A. (2016). Survival and
624 Evolution of a Large Multidrug Resistance Plasmid in New Clinical Bacterial Hosts.
625 Mol Biol Evol *33*, 2860-2873.
626 65. Silva, J.B., Storms, Z., and Sauvageau, D. (2016). Host receptors for bacteriophage
627 adsorption. Fems Microbiol Lett *363*.
628 66. San Millan, A., Pena-Miller, R., Toll-Riera, M., Halbert, Z.V., McLean, A.R., Cooper,
629 B.S., and MacLean, R.C. (2014). Positive selection and compensatory adaptation
630 interact to stabilize non-transmissible plasmids. Nat Commun *5*, 5208.
631 67. Harrison, E., Dytham, C., Hall, J.P., Guymer, D., Spiers, A.J., Paterson, S., and
632 Brockhurst, M.A. (2016). Rapid compensatory evolution promotes the survival of
633 conjugative plasmids. Mob Genet Elements *6*, e1179074.
634 68. Loftie-Eaton, W., Bashford, K., Quinn, H., Dong, K., Millstein, J., Hunter, S.,
635 Thomason, M.K., Merrikh, H., Ponciano, J.M., and Top, E.M. (2017). Compensatory
636 mutations improve general permissiveness to antibiotic resistance plasmids. Nat Ecol
637 Evol *1*, 1354-1363.
638 69. Harrison, E., Guymer, D., Spiers, A.J., Paterson, S., and Brockhurst, M.A. (2015).
639 Parallel compensatory evolution stabilizes plasmids across the parasitism-mutualism
640 continuum. Current biology : CB *25*, 2034-2039.
641 70. Yano, H., Wegrzyn, K., Loftie-Eaton, W., Johnson, J., Deckert, G.E., Rogers, L.M.,
642 Konieczny, I., and Top, E.M. (2016). Evolved plasmid-host interactions reduce
643 plasmid interference cost. Mol Microbiol *101*, 743-756.
644 71. Bottery, M.J., Wood, A.J., and Brockhurst, M.A. (2019). Temporal dynamics of
645 bacteria-plasmid coevolution under antibiotic selection. Isme J *13*, 559-562.
646 72. Bottery, M.J., Wood, A.J., and Brockhurst, M.A. (2017). Adaptive modulation of
647 antibiotic resistance through intragenomic coevolution. Nat Ecol Evol *1*, 1364-1369.
648 73. Porse, A., Schonning, K., Munck, C., and Sommer, M.O. (2016). Survival and
649 Evolution of a Large Multidrug Resistance Plasmid in New Clinical Bacterial Hosts.
650 Mol Biol Evol *33*, 2860-2873.
651 74. Turner, P.E., Williams, E.S., Okeke, C., Cooper, V.S., Duffy, S., and Wertz, J.E.
652 (2014). Antibiotic resistance correlates with transmission in plasmid evolution.
653 Evolution *68*, 3368-3380.
654 75. Bobay, L.M., Touchon, M., and Rocha, E.P.C. (2014). Pervasive domestication of
655 defective prophages by bacteria. P Natl Acad Sci USA *111*, 12127-12132.
656 76. Kottara, A., Hall, J.P.J., Harrison, E., and Brockhurst, M.A. (2018). Variable plasmid
657 fitness effects and mobile genetic element dynamics across Pseudomonas species.
658 FEMS microbiology ecology *94*.
659 77. Agrawal, N., Dasaradhi, P.V., Mohmmed, A., Malhotra, P., Bhatnagar, R.K., and
660 Mukherjee, S.K. (2003). RNA interference: biology, mechanism, and applications.
661 Microbiol Mol Biol Rev *67*, 657-685.
662 78. Lucchini, S., Rowley, G., Goldberg, M.D., Hurd, D., Harrison, M., and Hinton, J.C.
663 (2006). H-NS mediates the silencing of laterally acquired genes in bacteria. Plos
664 Pathog *2*, e81.
665 79. Marraffini, L.A., and Sontheimer, E.J. (2008). CRISPR interference limits horizontal
666 gene transfer in staphylococci by targeting DNA. Science *322*, 1843-1845.

667 80. Dupuis, M.E., Villion, M., Magadan, A.H., and Moineau, S. (2013). CRISPR-Cas and
668 restriction-modification systems are compatible and increase phage resistance. Nat
669 Commun *4*, 2087.
670 81. Oliveira, P.H., Touchon, M., and Rocha, E.P. (2016). Regulation of genetic flux
671 between bacteria by restriction-modification systems. Proc Natl Acad Sci U S A *113*,
672 5658-5663.
673 82. Palmer, K.L., and Gilmore, M.S. (2010). Multidrug-resistant enterococci lack
674 CRISPR-cas. MBio *1*.
675 83. Watson, B.N.J., Staals, R.H.J., and Fineran, P.C. (2018). CRISPR-Cas-Mediated
676 Phage Resistance Enhances Horizontal Gene Transfer by Transduction. MBio *9*.
677 84. Goldberg, G.W., Jiang, W., Bikard, D., and Marraffini, L.A. (2014). Conditional
678 tolerance of temperate phages via transcription-dependent CRISPR-Cas targeting.
679 Nature *514*, 633-637.
680 85. Faure, G., Makarova, K.S., and Koonin, E.V. (2019). CRISPR-Cas: Complex
681 Functional Networks and Multiple Roles beyond Adaptive Immunity. J Mol Biol *431*,
682 3-20.
683 86. Gophna, U., Kristensen, D.M., Wolf, Y.I., Popa, O., Drevet, C., and Koonin, E.V.
684 (2015). No evidence of inhibition of horizontal gene transfer by CRISPR-Cas on
685 evolutionary timescales. Isme J *9*, 2021-2027.
686 87. Gao, N.L., Chen, J., Lercher, M.J., and Chen, W.-H. (2018). Prokaryotic genome
687 expansion is facilitated by phages and plasmids but impaired by CRISPR. BioRxiv.
688 88. Doron, S., Melamed, S., Ofir, G., Leavitt, A., Lopatina, A., Keren, M., Amitai, G., and
689 Sorek, R. (2018). Systematic discovery of antiphage defense systems in the
690 microbial pangenome. Science *359*.
691 89. Thomas, C.M., and Nielsen, K.M. (2005). Mechanisms of, and barriers to, horizontal
692 gene transfer between bacteria. Nat Rev Microbiol *3*, 711-721.
693 90. Berngruber, T.W., Weissing, F.J., and Gandon, S. (2010). Inhibition of superinfection
694 and the evolution of viral latency. J Virol *84*, 10200-10208.
695 91. Faure, G., Shmakov, S.A., Yan, W.X., Cheng, D.R., Scott, D.A., Peters, J.E.,
696 Makarova, K.S., and Koonin, E.V. (2019). CRISPR-Cas in mobile genetic elements:
697 counter-defence and beyond. Nat Rev Microbiol.
698 92. Hartl, D.L., and Clark, A.G. (2007). Principles of population genetics, 4th Edition,
699 (Sunderland, Mass.: Sinauer Associates).
700 93. Charlesworth, B. (2009). Effective population size and patterns of molecular evolution
701 and variation. Nat Rev Genet *10*, 195-205.
702 94. San Millan, A., and MacLean, R.C. (2017). Fitness Costs of Plasmids: a Limit to
703 Plasmid Transmission. Microbiol Spectr *5*.
704 95. Vogwill, T., and MacLean, R.C. (2015). The genetic basis of the fitness costs of
705 antimicrobial resistance: a meta-analysis approach. Evol Appl *8*, 284-295.
706 96. Price, M.N., Wetmore, K.M., Waters, R.J., Callaghan, M., Ray, J., Liu, H., Kuehl, J.V.,
707 Melnyk, R.A., Lamson, J.S., Suh, Y., et al. (2018). Mutant phenotypes for thousands
708 of bacterial genes of unknown function. Nature *557*, 503-509.
709 97. van Opijnen, T., and Camilli, A. (2013). Transposon insertion sequencing: a new tool
710 for systems-level analysis of microorganisms. Nat Rev Microbiol *11*, 435-442.
711 98. Giaever, G., Chu, A.M., Ni, L., Connelly, C., Riles, L., Veronneau, S., Dow, S.,
712 Lucau-Danila, A., Anderson, K., Andre, B., et al. (2002). Functional profiling of the
713 Saccharomyces cerevisiae genome. Nature *418*, 387-391.
714 99. Sela, I., Wolf, Y.I., and Koonin, E.V. (2016). Theory of prokaryotic genome evolution.
715 P Natl Acad Sci USA *113*, 11399-11407.
716 100. Bobay, L.M., and Ochman, H. (2018). Factors driving effective population size and
717 pan-genome evolution in bacteria. Bmc Evolutionary Biology *18*.
718 101. Mira, A., Ochman, H., and Moran, N.A. (2001). Deletional bias and the evolution of
719 bacterial genomes. Trends Genet *17*, 589-596.

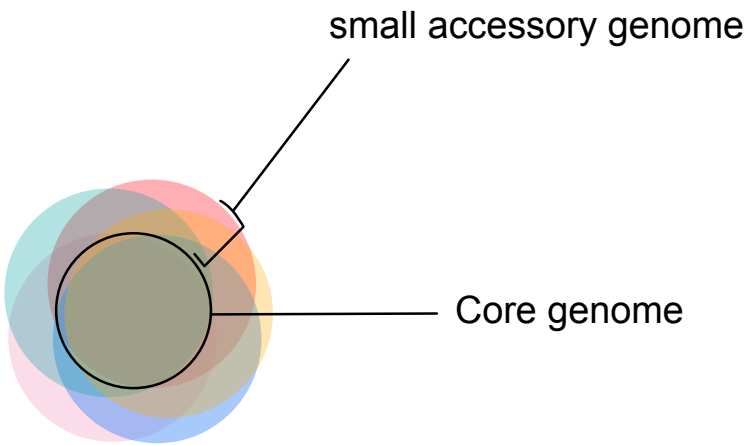
720 102. Gil, R., Sabater-Munoz, B., Latorre, A., Silva, F.J., and Moya, A. (2002). Extreme
721 genome reduction in Buchnera spp.: toward the minimal genome needed for
722 symbiotic life. Proc Natl Acad Sci U S A *99*, 4454-4458.
723 103. Veyrier, F.J., Dufort, A., and Behr, M.A. (2011). The rise and fall of the
724 Mycobacterium tuberculosis genome. Trends Microbiol *19*, 156-161.
725 104. Niehus, R., Mitri, S., Fletcher, A.G., and Foster, K.R. (2015). Migration and horizontal
726 gene transfer divide microbial genomes into multiple niches. Nature Communications
727 *6*.
728 105. Fraser, C., Alm, E.J., Polz, M.F., Spratt, B.G., and Hanage, W.P. (2009). The
729 bacterial species challenge: making sense of genetic and ecological diversity.
730 Science *323*, 741-746.
731 106. Andreani, N.A., Hesse, E., and Vos, M. (2017). Prokaryote genome fluidity is
732 dependent on effective population size. Isme J *11*, 1719-1721.
733 107. Collins, R.E., and Higgs, P.G. (2012). Testing the Infinitely Many Genes Model for
734 the Evolution of the Bacterial Core Genome and Pangenome. Mol Biol Evol *29*, 3413-
735 3425.
736 108. Baumdicker, F., Hess, W.R., and Pfaffelhuber, P. (2012). The Infinitely Many Genes
737 Model for the Distributed Genome of Bacteria. Genome Biol Evol *4*, 443-456.
738 109. Rocha, E.P.C. (2018). Neutral Theory, Microbial Practice: Challenges in Bacterial
739 Population Genetics. Mol Biol Evol *35*, 1338-1347.
740 110. May, R.M., and Anderson, R.M. (1983). Epidemiology and Genetics in the
741 Coevolution of Parasites and Hosts. P Roy Soc Lond a Mat *390*, 219-219.
742 111. Partridge, S.R., Kwong, S.M., Firth, N., and Jensen, S.O. (2018). Mobile Genetic
743 Elements Associated with Antimicrobial Resistance. Clin Microbiol Rev *31*.
744 112. Rozwandowicz, M., Brouwer, M.S.M., Fischer, J., Wagenaar, J.A., Gonzalez-Zorn,
745 B., Guerra, B., Mevius, D.J., and Hordijk, J. (2018). Plasmids carrying antimicrobial
746 resistance genes in Enterobacteriaceae. J Antimicrob Chemother *73*, 1121-1137.
747 113. Cordero, O.X., and Polz, M.F. (2014). Explaining microbial genomic diversity in light
748 of evolutionary ecology. Nat Rev Microbiol *12*, 263-273.
749 114. Gardner, A. (2009). Adaptation as organism design. Biology letters *5*, 861-864.
750 115. Karkman, A., Parnanen, K., and Larsson, D.G.J. (2019). Fecal pollution can explain
751 antibiotic resistance gene abundances in anthropogenically impacted environments.
752 Nat Commun *10*, 80.
753 116. Klumper, U., Riber, L., Dechesne, A., Sannazzarro, A., Hansen, L.H., Sorensen, S.J.,
754 and Smets, B.F. (2015). Broad host range plasmids can invade an unexpectedly
755 diverse fraction of a soil bacterial community. Isme J *9*, 934-945.
756 117. Datta, M.S., Sliwerska, E., Gore, J., Polz, M.F., and Cordero, O.X. (2016). Microbial
757 interactions lead to rapid micro-scale successions on model marine particles. Nat
758 Commun *7*, 11965.
759 118. Leibold, M.A. (2018). Metacommunity ecology, (Princeton, NJ: Princeton University
760 Press).
761 119. Makarova, K.S., Wolf, Y.I., Mekhedov, S.L., Mirkin, B.G., and Koonin, E.V. (2005).
762 Ancestral paralogs and pseudoparalogs and their role in the emergence of the
763 eukaryotic cell. Nucleic Acids Res *33*, 4626-4638.
764 120. Doolittle, R.F. (1995). The multiplicity of domains in proteins. Annu Rev Biochem *64*,
765 287-314.
766 121. Nevoigt, E., Fassbender, A., and Stahl, U. (2000). Cells of the yeast Saccharomyces
767 cerevisiae are transformable by DNA under non-artificial conditions. Yeast *16*, 1107-
768 1110.
769 122. Heinemann, J.A., and Sprague, G.F., Jr. (1989). Bacterial conjugative plasmids
770 mobilize DNA transfer between bacteria and yeast. Nature *340*, 205-209.
771 123. Soanes, D., and Richards, T.A. (2014). Horizontal gene transfer in eukaryotic plant
772 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

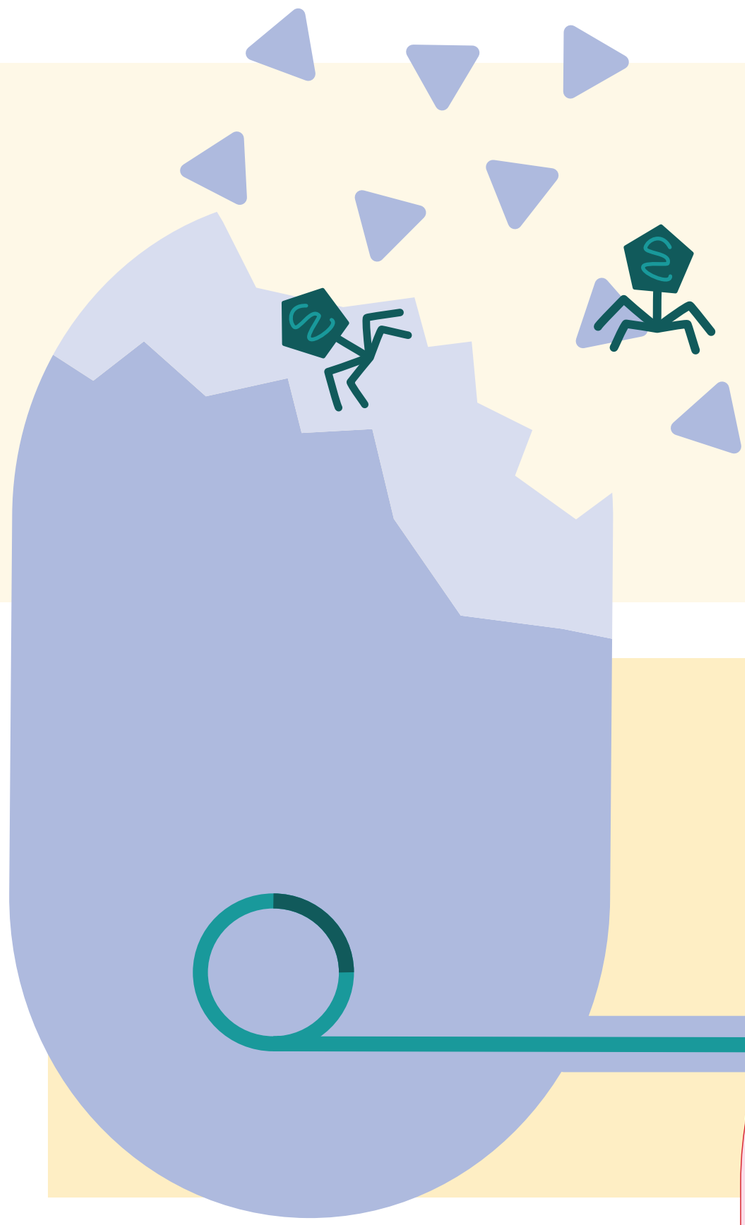



## Common among....

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