

The ecology and evolution of pangenomes

Brockhurst, Michael A; Harrison, Ellie; Hall, James P J; Richards, Thomas; McNally, Alan; MacLean, Craig

DOI:

[10.1016/j.cub.2019.08.012](https://doi.org/10.1016/j.cub.2019.08.012)

License:

Creative Commons: Attribution-NonCommercial-NoDerivs (CC BY-NC-ND)

Document Version

Peer reviewed version

Citation for published version (Harvard):

Brockhurst, MA, Harrison, E, Hall, JPJ, Richards, T, McNally, A & MacLean, C 2019, 'The ecology and evolution of pangenomes', *Current Biology*, vol. 29, no. 20, pp. R1094-R1103. <https://doi.org/10.1016/j.cub.2019.08.012>

[Link to publication on Research at Birmingham portal](#)

General rights

Unless a licence is specified above, all rights (including copyright and moral rights) in this document are retained by the authors and/or the copyright holders. The express permission of the copyright holder must be obtained for any use of this material other than for purposes permitted by law.

- Users may freely distribute the URL that is used to identify this publication.
- Users may download and/or print one copy of the publication from the University of Birmingham research portal for the purpose of private study or non-commercial research.
- User may use extracts from the document in line with the concept of 'fair dealing' under the Copyright, Designs and Patents Act 1988 (?)
- Users may not further distribute the material nor use it for the purposes of commercial gain.

Where a licence is displayed above, please note the terms and conditions of the licence govern your use of this document.

When citing, please reference the published version.

Take down policy

While the University of Birmingham exercises care and attention in making items available there are rare occasions when an item has been uploaded in error or has been deemed to be commercially or otherwise sensitive.

If you believe that this is the case for this document, please contact UBIRA@lists.bham.ac.uk providing details and we will remove access to the work immediately and investigate.

1 THE ECOLOGY AND EVOLUTION OF PANGENOMES

2 Michael A Brockhurst^{1*}, Ellie Harrison¹, James PJ Hall², Thomas Richards³, Alan McNally⁴,
3 Craig MacLean⁵

4

5 * Corresponding Author

6

7 1. Department of Animal and Plant Sciences, University of Sheffield, Sheffield UK

8 2. Institute for Integrative Biology, University of Liverpool, Liverpool UK

9 3. Biosciences, University of Exeter, Exeter UK

10 4. Institute of Microbiology and Infection, University of Birmingham, Birmingham UK

11 5. Department of Zoology, University of Oxford, Oxford UK

12

13 **Abstract**

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

25

26 **The pangenome concept**

27 The pangenome describes all the genes present in a species and can be subdivided into
28 those shared by all members of a species—the core genes—and those present in only some
29 members of a species—the accessory genes [1] (Figure 1). The pangenome concept
30 emerged from early comparative studies of bacterial genomes. Comparison of a pathogenic
31 *Escherichia coli* O157 strain with its non-pathogenic relative *E. coli* K12 MG1655, showed

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

53 The current challenge is to move beyond this phenomenological description of pangenomes
54 to forge an understanding of the mechanisms and processes that determine their structure.
55 A genome sequence is a snapshot of a strain in time. Some of the genes and mutations in
56 that snapshot share a long history and are destined to remain associated, while other
57 members are transient: recent acquisitions, or in the process of leaving. How do we
58 distinguish between these categories? If a genome is a family photograph, how do we

59 distinguish real members from the photobombers? A starting point is to understand the
60 processes and mechanisms that promote or prevent gene gain and loss, and thereby cause
61 dynamic flux in the content of the pangenome. Gene gain by a lineage in the context of the
62 pan-genome can be conceptually separated into two distinct processes, operating on
63 different timescales and affected by different environmental drivers. The first describes the
64 specific gene acquisition event, which occurs at the level of individual cells and is effectively
65 instantaneous, while the second represents the stable assimilation of acquired genes within
66 populations and their non-random elimination from a lineage, and is on-going, with effects
67 emerging over a longer period and in different ways in different environments. In this review,
68 we first outline the molecular, ecological and evolutionary drivers of gene gain and loss
69 which mediate changes in the composition of the pangenome, and then discuss how
70 evolutionary theory can be applied to understand the structure of pangenomes.

71

72 **Drivers and barriers of gene gain and loss**

73 Gene acquisition introduces variation, and thus provides the raw material upon which
74 selection can subsequently act [7]. Various mechanisms actively facilitate the movement of
75 genetic material across membranes, and these are particularly well-described in prokaryotes
76 but there is evidence that equivalent mechanisms may exist in model eukaryotes such as
77 yeast. In recent decades, the canonical processes — conjugation, transduction, and
78 transformation — have been joined by more recently-characterised phenomena, including
79 nanotubes [8] and vesicles [9]. These varied mechanisms of gene exchange offer the
80 potential for gene acquisition, but the likelihood of its occurrence depends on a range of
81 ecological, mechanistic and evolutionary factors, explored in this section (summarised in
82 Figure 2).

83

84 *Ecological opportunity for HGT*

85 The proximal environmental triggers activating expression of gene exchange machinery vary
86 between systems and with different species, but some common themes can be identified.
87 One of these is stress. For example, the SOS response to DNA damage, triggered by some
88 antibiotics, reactive oxygen, and UV radiation, activates transfer of the *Vibrio cholerae* STX
89 element [10], causes integron rearrangement [11], and activates integrated bacteriophage
90 [12]. Transposons in *E. coli* become active under nutritional stress [13], plasmid conjugation
91 rates are increased in response to host inflammation in mammalian gut [14], and starvation
92 conditions activate natural competence [15]. However, different stress responses can lead to
93 divergent effects in different species [16], and donors, recipients, and mobile genetic
94 elements may each respond to different cues. For example, some mobile genetic elements,
95 such as the 'pheromone-inducible' conjugative plasmids of *Enterococcus*, have evolved
96 mechanisms to sense the presence of recipients [17], and transformation is induced by
97 quorum sensing and by specific nutrients in some species of *Vibrio* [18].

98 Ecology appears to be a principal determinant of gene-sharing [19] suggesting that the
99 transfer of genes is to some extent limited by ecological opportunity. Several important gene
100 transfer mechanisms including conjugation and nanotubes require close physical proximity
101 and thus HGT is probabilistically likely to be most efficient between immediate neighbours
102 [20]. Consequently, the size of the gene pool from which a species can draw will be
103 dependent on the diversity of environments they occupy as well as the community diversity
104 these contain. Correspondingly, networks of gene sharing have shown that co-occurrence of
105 species in a habitat increases the probability of gene sharing [21-24]. Niche specialists likely
106 to exist in stable environments with very low diversity, such as endosymbionts [23], have
107 more closed pan-genomes than those that exist in diverse communities and more variable
108 environments.

109 Among symbionts and pathogens with low rates of gene gain through HGT, variation in gene
110 loss among lineages can be the primary cause of diversity among clonal lineages, and can
111 lead to large phenotypic differences [25]. Whereas gene loss can be positively selected in

112 large populations with efficient selection, in intracellular symbionts and pathogens with low
113 effective population size gene loss is more likely to be a result of relaxed selection and drift
114 [26]. How the balance of gene gain and loss contributes to the formation of a pangenome is
115 well-illustrated by *Yersinia enterocolitica*. The species is composed of five phylogenetically
116 distinct groups, four of which are pathogenic to humans and have emerged from a non-
117 pathogenic ancestor, driven by a single acquisition of a large virulence plasmid [27].
118 Following plasmid acquisition, the splits between the four pathogenic groups are delineated
119 at a pangenome level by differential loss genes present in the ancestor, alongside HGT
120 leading to switches in serotype [28].

121

122 *Mechanistic drivers and barriers of HGT*

123 Once acquired there are significant barriers to the maintenance of novel genetic material
124 which shape the patterns of gene sharing among species. Newly acquired DNA must
125 replicate to ensure it is passed to daughter cells, either by carrying with it replication
126 machinery compatible with that of the host (in the case of plasmids) or by integrating into a
127 resident replicon. Integration can occur through general recipient-encoded processes such
128 as homologous recombination which is dependent on regions of sequence homology [29,
129 30] or by the activity of entities such as transposons, integrons, and insertion sequences,
130 which can facilitate capture of incoming DNA (e.g., [31]). Finally, genes must be able to
131 function in the host in order to have a phenotypic effect subject to selection, which is
132 dependent on recognition of promoters allowing for gene expression [32], and comparable
133 GC content, codon usage and compatible genetic codes allowing for efficient translation [33],
134 and in the case of DNA transfer between eukaryotic genomes effective splicing of introns. As
135 a general principle, many of these processes become more challenging across larger
136 genetic distances [34]. Correspondingly gene sharing has been shown to be most common
137 between closer relatives [24].

138 Mechanistic limitations are also likely to define the types of genes that are more readily
139 shared, and therefore more likely to contribute to the accessory genome. Incoming DNA can
140 disrupt cellular processes leading to severe fitness costs, and these genes are likely to be
141 rapidly lost from the population by purifying selection. Genes encoding core cellular
142 functions, such as those associated with transcription and translation, are highly toxic when
143 expressed in foreign hosts [32, 35] and poorly represented among horizontally transferred
144 genes [36, 37]. This strong incompatibility may be associated less with function *per se*,
145 rather than the number of protein-protein interactions which the encoded protein engages in.
146 Genes embedded within more complex interaction networks are more disruptive and less
147 likely to maintain the necessary functional interaction network when transferred, a
148 phenomenon termed the complexity hypothesis [38, 39]. MGEs themselves are often
149 associated with significant fitness costs that are caused by a range of factors, including the
150 biosynthetic cost of maintaining and expressing additional DNA, toxic gene products, and
151 epistasis between chromosomal and MGE-encoded genes [40]. This disruptive effect of
152 HGT is not surprising from an evolutionary perspective: HGT brings together genes that
153 have fundamentally different evolutionary histories, and there is no a priori reason to expect
154 that these genes should function together harmoniously [41].

155

156 *Evolutionary conflict and collaboration in the pangenome*

157 Many of the mechanisms for horizontal gene transfer are encoded by infectious MGEs such
158 as viruses, plasmids, and transposable elements. Therefore, pangenomes are composites of
159 the host chromosome(s) together with MGEs that may be shared with other species. MGEs
160 encode accessory genes that may represent adaptive additions to the pangenome (e.g. by
161 providing a new ecological function or access to an otherwise inaccessible niche), but also
162 encode genes for MGE-related functions such as replication and transmission, as well as
163 many genes of unknown function. As semi-autonomous evolving entities we should expect
164 MGEs to maximise their own fitness through both vertical and horizontal transmission [42].

165 Encoding beneficial accessory genes can enhance MGE fitness through enhanced vertical
166 transmission. However, being beneficial is not necessary for MGE success. Many
167 environmental plasmids do not encode any obvious accessory genes [43] and are therefore
168 likely to be genetic parasites. Experimental studies show that high rates of horizontal
169 transmission through conjugation can maintain costly resistance plasmids in the absence of
170 positive selection [44-46], and non-beneficial plasmids can invade biofilm populations [47,
171 48]. Indeed, experiments with antibiotic resistance [49] and mercury detoxification [46]
172 plasmids have shown that positive selection for these functions limits the horizontal transfer
173 of these resistance genes by reducing the availability of recipient cells [46, 49]. Although, in
174 the long run, purely infectious elements would be expected to become increasingly efficient
175 parasites by shedding their accessory genes, mobile genetic elements that persist through
176 horizontal transmission are likely to be especially prone to mediating gene exchange [50].
177 Higher rates of horizontal transmission are likely to expose these MGEs to a wider diversity
178 of genomic environments, offering greater opportunity for other MGEs (e.g., transposons) to
179 integrate and hitch a ride.

180 The predominance of gene exchange mediated by MGEs means that this form of gene
181 sharing is, at least partially, constrained by the host range of MGEs. Phages are believed to
182 have relatively narrow host ranges, which are often limited to within a species or genus [51,
183 52]. Plasmid host ranges can be broader, and are dependent on the diversity of replication
184 genes required for stable maintenance in different host taxa [53]. Correspondingly, plasmids
185 appear to be more important mediators of gene exchange across larger genetic distances
186 [54]. However, interactions between MGEs allow smaller, simpler elements to escape these
187 restrictions. Transposable elements like transposons, which are themselves unable to
188 transfer between cells, can hitch a ride on a conjugative plasmid, as has been observed for
189 plasmid-encoded antibiotic resistances in hospital outbreaks of Enterobacteriaceae [55, 56].
190 Further transfer of transposons between plasmids with different host ranges then expands
191 the range of potential hosts accessible to these transposon-encoded genes. Plasmids too

192 can be composite mosaics of other elements, including other plasmids, broadening the
193 range of hosts in which they can replicate, while transposons can become nested within one
194 another, increasing opportunities for spread [57]. A consequence of the self-interested
195 activity of MGEs for genome evolution is that 'selfish' genes spread between lineages
196 alongside the MGE-encoded accessory functions that enhance host fitness or niche
197 adaptation. Indeed, plasmid, phage, and transposon-encoded functions are usually highly
198 represented in the pangenome and in comparative studies of horizontal gene transfer [5, 58].

199 Because they can replicate by both vertical and horizontal transmission, MGEs can have
200 fitness interests that do not necessarily align with those of other parts of the (vertically-
201 inherited) genome. These 'divided loyalties' manifest in the fitness costs associated with
202 MGE acquisition and horizontal transmission, and result in intragenomic conflict. For
203 example, while conjugation provides an efficient mechanism for plasmids to transfer
204 between bacteria, the expression of conjugative machinery imposes a biosynthetic fitness
205 cost on the donor cell [59], and leaves the donor cell open to predation by pilus-targeting
206 phage [60]. Resolution of host-MGE conflict frequently requires compensatory evolution to
207 reduce the fitness costs of the newly acquired genes [42], and is promoted by positive
208 selection for MGE-encoded functions since this increases the population size and mutation
209 supply for MGE-carriers [61, 62]. Diverse compensatory mechanisms have been identified to
210 stabilise plasmids, but two common routes are mutations affecting host gene regulatory
211 networks [63, 64] or plasmid replication [41, 65]. By stabilising MGEs within bacterial
212 lineages, compensatory evolution can set the stage for more extensive coevolution between
213 the MGE and chromosome, driving reciprocal adaptations and counter-adaptations [42]. For
214 example, bacteria-plasmid coevolution rapidly led to the emergence of co-dependence of
215 chromosomal and plasmid replicons under antibiotic selection, together providing high-level
216 resistance but separately providing inadequate resistance to persist in the environment they
217 evolved in [66, 67]. Compensation and coevolution can, in turn, drive the complete
218 domestication of MGEs and their integration into a more exclusively vertical mode of

219 replication. In practice, domestication involves downregulation, inactivation, or loss of the
220 machinery involved in horizontal transmission, through gene deletion [68, 69]. Bacterial
221 genomes contain numerous prophages, some of which are incapable of horizontal
222 transmission and now serve their bacterial hosts as anti-competitor toxins [70]. Alternatively,
223 recombination can relocate mobile genes to non-mobile parts of the genome, e.g. capture of
224 resistance genes from plasmids, a process rapid enough to be readily observable in the
225 laboratory [45, 64, 71]. In so doing, the signatures of gene acquisition are gradually lost from
226 the genome sequence, potentially explaining why many accessory genes in pangenomes
227 are no longer obviously associated with MGEs.

228

229 *Resisting HGT*

230 Due to the potential for conflict between MGEs and the host chromosome, immunity systems
231 which actively target incoming foreign DNA are widespread across eukaryotes and
232 prokaryotes. Systems exist in both eukaryotes (e.g. RNAi [72]) and prokaryotes (e.g. H-NS
233 [73]) to silence gene expression from foreign DNA. In prokaryotes CRISPR-Cas systems
234 and restriction-modification (R-M) systems target novel DNA for degradation, and can be an
235 effective defence against MGEs, consequently reducing HGT [74, 75]. A comparative
236 analysis of 79 prokaryote genomes show that R-M systems structure gene sharing by
237 favouring exchanges between genomes with similar R-M systems [76]. The relationship
238 between HGT and CRISPR-Cas systems appears more complex: There are well-described
239 cases where CRISPR-Cas systems are negatively associated with MGE carriage within
240 species [77], but CRISPR-Cas have also been shown to promote HGT in some cases [78].
241 Type-III CRISPR-Cas systems target actively transcribed DNA via spacers derived from
242 RNA transcripts [79] and may therefore be more effective against phages and plasmids than
243 DNA acquired by transformation [80]. Over broader taxonomic scales, however, the
244 correlation between CRISPR-Cas systems and the rate of HGT is less clear and deserves
245 further study [81, 82]. It is likely that other similar mechanisms will continue to be discovered

246 [83]. Resistance mechanisms protecting cells against incoming DNA can also be encoded by
247 MGEs themselves, highlighting how conflict between MGE could act to limit HGT. Both
248 plasmids and phages defend their host cells against super-infection through self-exclusion
249 mechanisms [84, 85] and can encode their own CRISPR-Cas systems with spacer
250 sequences targeting other MGEs [86].

251

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

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

263

264 *A population genetic approach to the pangenome*

265 Evolutionary biologists have developed a mature body of population genetic theory to
266 understand how mutation, selection and genetic drift interact to shape patterns of genetic
267 variation [87]. A key insight from population genetic theory is that the efficacy of natural
268 selection is critically dependent on population size [88]: in species with a low effective
269 population size, selection is weak relative to the genetic drift and evolution is dominated by
270 the stochastic spread of weakly deleterious mutations. In contrast, natural selection is a
271 strong force relative to genetic drift in species with a high effective population size. Under

272 these conditions, selection prevents the spread of weakly deleterious mutations and drives
273 selective sweeps of beneficial mutations. Like spontaneous mutation, both gene acquisition
274 [34, 40, 89, 90] and loss [91-93] tend to reduce fitness. Therefore, selection should shape
275 patterns of gene gain and loss in species with high N_e , whereas selection will have reduced
276 potency in species with low N_e and therefore the genome evolution and the extent and
277 composition of the pangenome in such species will be susceptible to underlying rates of
278 gene gain and loss.

279 A number of studies have shown that average genome size is large in bacterial species with
280 a large effective population size [94, 95]. The simplest explanation for this correlation is that
281 drift allows the accumulation of weakly deleterious deletions in species with low N_e .
282 Therefore, gene loss occurs at a greater rate than gene acquisition in bacterial genomes [96]
283 of species with smaller N_e , driving a trend of genome reduction. For example, genomic
284 degeneration is commonly observed in species that undergo recurrent population
285 bottlenecks during transmission, such as endosymbiotic bacteria [97] and intracellular
286 pathogens [98]. Many genes in bacterial genomes only provide a fitness benefit under very
287 specific environmental conditions [91], and effective selection for marginally beneficial genes
288 acquired by HGT in species with high N_e is also likely to contribute to the positive correlation
289 between N_e and genome size. Simply put, because species with large N_e are likely to
290 occupy wider environment profiles, they are also likely to be under a wider diversity of
291 environmental conditions driving selection for gene diversity and therefore larger genome
292 sizes (Figure 1). As such species with high N_e also have large pangenomes [95, 99], and
293 [99] argue that this correlation is evidence that the pangenome is adaptive. The concept of
294 population structure is key to this argument: in species with low levels of population
295 structure, adaptive gene acquisition and loss events will sweep to fixation, and these will
296 therefore not contribute to the pangenome. Population subdivision provides the opportunity
297 for selection to contribute to increasing the pangenome size of a species because selective
298 sweeps of locally adaptive gene gain and loss events will affect the pangenome size [100].

299 Other studies using population genetics have questioned the role of selection in shaping the
300 pangenome. Comparing levels of synonymous nucleotide diversity, a surrogate measure of
301 N_e , with a measure pangenome fluidity showed a positive correlation between N_e and
302 pangenome fluidity, that could arise because genetic drift leads to the loss of effectively
303 neutral accessory genes in species with low N_e [101]. Further support for this idea comes
304 from comparing the observed distribution of gene frequencies in the pangenome with an
305 expected distribution generated by a neutral model. This approach, inspired by the infinite
306 alleles model, assumes that bacteria gain genes from an infinite pool of horizontally
307 transferred genes and subsequently lose these genes through drift [102, 103]. Accessory
308 genes show a distribution that is close to the expectations of a neutral model for widely
309 distributed marine bacteria, but with some deviations that are consistent with selection
310 shaping the pangenome [103]. It is unclear, however, that currently available genomic data
311 provide the necessary breadth and depth of ecological sampling to adequately test these
312 models.

313

314 *The limits of a population genetic approach*

315 Population genetics theory provides some simple guiding principles for understanding the
316 pangenome, but there are also potential difficulties with applying these models to understand
317 the pangenome [104]. For example, classical population genetic tests for selection rely on
318 comparing observed patterns of genetic polymorphisms and divergence with expected
319 patterns from a neutral model where evolution is driven by mutation and drift, but not
320 selection. Neutral models in population genetics assume that mutations at different sites in
321 the genome are not linked. This is a justifiable assumption in eukaryotic species with
322 obligate sexual reproduction, but the pangenome changes through the gain and loss of
323 blocks of genes, for example because they are all encoded on a mobile genetic element. An
324 important consequence of this is that strong selection for one gene (e.g. an antibiotic
325 resistance gene) can lead to the spread of linked mildly deleterious genes by co-selection, if

326 there is a net fitness benefit of the MGE. Similarly, genes that are linked to addiction
327 systems, such as toxin-antitoxin systems, can be maintained in populations by the toxic
328 effects of MGE loss. In a broader perspective, the strong linkage disequilibrium observed in
329 clonal bacterial species means that there might be no effectively neutral variation [104].

330 A second important difficulty is that population genetic models ignore the evolutionary
331 conflicts of interest that can occur between accessory and core genes in the same genome.

332 A key concept from evolutionary ecology is that trade-offs exist between the efficacy of
333 vertical and horizontal transmission [105], preventing the evolution of elements that are to
334 provide a big benefit to their host and transfer efficiently between hosts. Trade-offs may also
335 limit the ability of MGEs to maximize the fitness benefit that they provide to different hosts,
336 further limiting the benefits that hosts gain from acquiring MGEs [67]. All else being equal,
337 we would therefore expect that MGEs with high mobility, such as broad-host range
338 conjugative plasmids and lysogenic phage, to impose greater fitness costs than genetic
339 elements with a low mobility, such as non-transmissible plasmids and defective prophage.
340 This logic is somewhat counter-intuitive, because many of the most obviously adaptive
341 genes in the pangenome, such as antibiotic resistance genes, are often found on MGEs with
342 high mobility [106, 107], but these adaptive genes may be 'rubbies in the rubbish' from the
343 perspective of their bacterial hosts.

344

345 **Perspective**

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

- 352 • Adaptation is the "process of optimisation of the phenotype under the action of natural
353 selection" [108]. As a pangenome emerges as an analytical result from comparing
354 multiple genomes, we must take care when specifying what adaptation means in this
355 context, i.e. who or what is being optimised. While a pangenome *can* contain adaptive
356 genes that are transferred between species, the pangenome does not evolve *for the*
357 *purposes of* maintaining a pool of niche-adaptive genes. Instead, its contents are defined
358 by selection occurring at lower organisational levels: the individual bacterial lineage that
359 has acquired locally-beneficial genes, and the persistent mobile genetic element. Neither
360 does a broadly adaptive pangenome imply that the accessory genes in a given genome
361 are beneficial to that strain. Recent migration [109] or gene acquisition can result in a
362 strain carrying neutral or deleterious genes which have not yet been lost. Finally, if the
363 pangenome is defined as the sum-total of all genes in a species, increased sequencing
364 resolution will increasingly capture rare or transient events and thus inflate the size of the
365 pangenome. Enhanced biological insight into the gene function, as well as bioinformatic
366 tools that help us distinguish between transient associations and longer-term
367 partnerships, will guard us from incorrectly inferring adaptation in such instances.
- 368 • The rate of horizontal gene transfer is key to both the population genetic and eco-evo
369 perspectives on the pangenome, but our knowledge of rate of HGT in the wild remains
370 very limited. It might be possible to measure these rate by using statistical methods to
371 infer rates of HGT from genomic data, and experimental methods that allow the spread
372 of genes to be measured under natural communities in real time using for example
373 microcosm experiments [50, 110].
- 374 • Microbial genomes are being sequenced at an incredible rate, but it is very challenging
375 to understand sequence data in a population genetics context, there are often huge
376 sampling biases in microbial sequence datasets (intensive sampling of clinical outbreaks
377 is the most extreme example). Given the vast population size of microbes, we will only
378 ever be able to achieve very sparse sampling of microbial genomes, even with the most

379 ambitious sequencing projects. We therefore need to develop approaches to identify and
380 sample ecologically coherent microbial populations [111]. For example, it is clear that
381 some microbial populations are structured at an incredibly fine scale, such as individual
382 particles of detritus [112], and this structuring can play a key role in the evolution of the
383 pangenome [100]. For example, comparing a small number of bacterial genomes
384 sampled from many niches is likely to produce an abundance of rare accessory genes,
385 but these could either represent adaptive accessory genes that are locally abundant but
386 globally rare, or deleterious accessory genes that are both locally and globally rare. One
387 key technological development that may help with this problem is to move from
388 sequencing the genomes of bacterial isolates to single-cell sequencing of bacteria from
389 environmental samples.

390 • The neutral theory of molecular evolution has been so useful in revealing the action of
391 natural selection because it makes quantitative and falsifiable predictions that be tested
392 by comparing datasets. Given the complexity of forces shaping the pangenome it may be
393 necessary to look outside of genetics for potential approaches: Pangenomes share many
394 characteristics with metacommunities, most notably the idea that entities (genes or
395 species) are sampled from a pool to form discrete sets (genomes or communities) that
396 share biological cohesiveness (pangenome or metacommunity). Metacommunity ecology
397 has a well-developed body of theory to understand how communities are assembled and
398 structured [113], which may help to unravel the processes causing the structure of
399 pangenomes.

400

401

402 **BOX 1: Do eukaryotes have pangenomes?** The existence of pangenomes in eukaryotes is
403 controversial [6]. What is evident is that, unlike the situation in prokaryotes, genome
404 evolution in eukaryotes is dominated by processes other than HGT, including sexual
405 recombination and gene duplication [114] often combined with domain reshuffling [115].
406 Nevertheless, HGT can and does occur: For example, *Saccharomyces* undergoes
407 transformation under starvation conditions [116] and can receive DNA by conjugation from
408 bacteria [117], although HGT from prokaryotes contributes just 0.5% of the gene repertoire
409 of *Saccharomyces* (reviewed in [118]). Additionally, a range of other mechanisms introduce
410 genetic material into eukaryotic cytoplasm offering the potential for HGT, including: viral
411 vectors [119], integration of viral fragments [120], RNA exchange, trophic interactions
412 through phagocytosis of prey cells [121], and anastomosis of cell structures [118, 122]. The
413 role of HGT in accessory genome variation is unclear, but likely to be far less important
414 than in prokaryotes and a relatively minor contributor compared to other factors like strain
415 level duplication [123] and differential gene loss. Pangenome studies in eukaryotes are
416 challenging due to their more complex genome architectures and a lack of replete genome-
417 level sampling. Analyses of model fungi suggest core genome fractions of between 80-90%
418 [123], whilst in the marine alga *Emiliania huxleyi*, 17% of genes present in the assembled
419 genome of the model strain CCMP1516 were absent in four other strains, indicating a
420 putative accessory genome [124]. Consistent with the complexity of eukaryotic genome
421 architecture, distinct dispensable or supernumerary chromosomes systems are observed
422 in some fungi that show signs of HGT derivation, operate to carry an accessory genome,
423 and define the niche and host range of the recipient lineage [125-127]. Therefore, while
424 the existing studies suggest that the pangenome concept is valid for eukaryotic microbes,
425 the extent of accessory genome variation is likely to be far lower than in prokaryotes: ~10-
426 15% of genes in eukaryotes compared to up to ~65% in some prokaryotes.

427

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

432

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

437

438 **Acknowledgements**

439 This work was funded by grants from BBSRC (BB/R014884/1 to MAB, JPJH, EH;
440 BB/R006253/1 to MAB; BB/R006261/1 to AM), NERC (NE/P017584/1 to EH;
441 NE/R008825/1 to MAB, JPJH, EH; NE/S000771/1 to MAB) and the Wellcome Trust
442 (106918/Z/15/Z to CM).

443

444

445

446 **Cited references**

447

- 448 1. Tettelin, H., Massignani, V., Cieslewicz, M.J., Donati, C., Medini, D., Ward, N.L.,
449 Angiuoli, S.V., Crabtree, J., Jones, A.L., Durkin, A.S., et al. (2005). Genome analysis
450 of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the
451 microbial "pan-genome". *Proc Natl Acad Sci U S A* *102*, 13950-13955.
- 452 2. Perna, N.T., Plunkett, G., 3rd, Burland, V., Mau, B., Glasner, J.D., Rose, D.J.,
453 Mayhew, G.F., Evans, P.S., Gregor, J., Kirkpatrick, H.A., et al. (2001). Genome
454 sequence of enterohaemorrhagic *Escherichia coli* O157:H7. *Nature* *409*, 529-533.
- 455 3. Welch, R.A., Burland, V., Plunkett, G., 3rd, Redford, P., Roesch, P., Rasko, D.,
456 Buckles, E.L., Liou, S.R., Boutin, A., Hackett, J., et al. (2002). Extensive mosaic
457 structure revealed by the complete genome sequence of uropathogenic *Escherichia*
458 *coli*. *Proc Natl Acad Sci U S A* *99*, 17020-17024.
- 459 4. Dobrindt, U., Hochhut, B., Hentschel, U., and Hacker, J. (2004). Genomic islands in
460 pathogenic and environmental microorganisms. *Nat Rev Microbiol* *2*, 414-424.
- 461 5. McInerney, J.O., McNally, A., and O'Connell, M.J. (2017). Why prokaryotes have
462 pangenomes. *Nat Microbiol* *2*, 17040.
- 463 6. Martin, W.F. (2017). Too Much Eukaryote LGT. *Bioessays* *39*.
- 464 7. Vos, M., Hesselman, M.C., Te Beek, T.A., van Passel, M.W.J., and Eyre-Walker, A.
465 (2015). Rates of Lateral Gene Transfer in Prokaryotes: High but Why? *Trends in*
466 *microbiology* *23*, 598-605.
- 467 8. Dubey, G.P., and Ben-Yehuda, S. (2011). Intercellular nanotubes mediate bacterial
468 communication. *Cell* *144*, 590-600.
- 469 9. Fulsundar, S., Harms, K., Flaten, G.E., Johnsen, P.J., Chopade, B.A., and Nielsen,
470 K.M. (2014). Gene transfer potential of outer membrane vesicles of *Acinetobacter*
471 *baylyi* and effects of stress on vesiculation. *Appl Environ Microbiol* *80*, 3469-3483.
- 472 10. Beaber, J.W., and Waldor, M.K. (2004). Identification of operators and promoters that
473 control SXT conjugative transfer. *J Bacteriol* *186*, 5945-5949.
- 474 11. Guerin, E., Cambray, G., Sanchez-Alberola, N., Campoy, S., Erill, I., Da Re, S.,
475 Gonzalez-Zorn, B., Barbe, J., Ploy, M.C., and Mazel, D. (2009). The SOS response
476 controls integron recombination. *Science* *324*, 1034.
- 477 12. Nanda, A.M., Thormann, K., and Frunzke, J. (2015). Impact of spontaneous
478 prophage induction on the fitness of bacterial populations and host-microbe
479 interactions. *J Bacteriol* *197*, 410-419.
- 480 13. Twiss, E., Coros, A.M., Tavakoli, N.P., and Derbyshire, K.M. (2005). Transposition is
481 modulated by a diverse set of host factors in *Escherichia coli* and is stimulated by
482 nutritional stress. *Mol Microbiol* *57*, 1593-1607.
- 483 14. Stecher, B., Denzler, R., Maier, L., Bernet, F., Sanders, M.J., Pickard, D.J., Barthel,
484 M., Westendorf, A.M., Krogfelt, K.A., Walker, A.W., et al. (2012). Gut inflammation
485 can boost horizontal gene transfer between pathogenic and commensal
486 *Enterobacteriaceae*. *Proc Natl Acad Sci U S A* *109*, 1269-1274.
- 487 15. Blokesch, M. (2016). Natural competence for transformation. *Current biology : CB* *26*,
488 R1126-R1130.
- 489 16. Johnston, C., Martin, B., Fichant, G., Polard, P., and Claverys, J.P. (2014). Bacterial
490 transformation: distribution, shared mechanisms and divergent control. *Nat Rev*
491 *Microbiol* *12*, 181-196.
- 492 17. Koraimann, G., and Wagner, M.A. (2014). Social behavior and decision making in
493 bacterial conjugation. *Front Cell Infect Microbiol* *4*, 54.
- 494 18. Seitz, P., and Blokesch, M. (2013). DNA-uptake machinery of naturally competent
495 *Vibrio cholerae*. *Proc Natl Acad Sci U S A* *110*, 17987-17992.
- 496 19. Smillie, C.S., Smith, M.B., Friedman, J., Cordero, O.X., David, L.A., and Alm, E.J.
497 (2011). Ecology drives a global network of gene exchange connecting the human
498 microbiome. *Nature* *480*, 241-244.

- 499 20. Babic, A., Berkmen, M.B., Lee, C.A., and Grossman, A.D. (2011). Efficient gene
500 transfer in bacterial cell chains. *MBio* 2.
- 501 21. Hooper, S.D., Mavromatis, K., and Kyrpides, N.C. (2009). Microbial co-habitation and
502 lateral gene transfer: what transposases can tell us. *Genome Biol* 10, R45.
- 503 22. Chaffron, S., Rehrauer, H., Pernthaler, J., and von Mering, C. (2010). A global
504 network of coexisting microbes from environmental and whole-genome sequence
505 data. *Genome Res* 20, 947-959.
- 506 23. Kloesges, T., Popa, O., Martin, W., and Dagan, T. (2011). Networks of gene sharing
507 among 329 proteobacterial genomes reveal differences in lateral gene transfer
508 frequency at different phylogenetic depths. *Mol Biol Evol* 28, 1057-1074.
- 509 24. Popa, O., and Dagan, T. (2011). Trends and barriers to lateral gene transfer in
510 prokaryotes. *Curr Opin Microbiol* 14, 615-623.
- 511 25. Bolotin, E., and Hershberg, R. (2015). Gene Loss Dominates As a Source of Genetic
512 Variation within Clonal Pathogenic Bacterial Species. *Genome Biol Evol* 7, 2173-
513 2187.
- 514 26. McNally, A., Thomson, N.R., Reuter, S., and Wren, B.W. (2016). 'Add, stir and
515 reduce': *Yersinia* spp. as model bacteria for pathogen evolution. *Nat Rev Microbiol*
516 14, 177-190.
- 517 27. Reuter, S., Connor, T.R., Barquist, L., Walker, D., Feltwell, T., Harris, S.R., Fookes,
518 M., Hall, M.E., Petty, N.K., Fuchs, T.M., et al. (2014). Parallel independent evolution
519 of pathogenicity within the genus *Yersinia*. *Proc Natl Acad Sci U S A* 111, 6768-6773.
- 520 28. Reuter, S., Corander, J., de Been, M., Harris, S., Cheng, L., Hall, M., Thomson, N.R.,
521 and McNally, A. (2015). Directional gene flow and ecological separation in *Yersinia*
522 *enterocolitica*. *Microb Genom* 1, e000030.
- 523 29. Majewski, J., and Cohan, F.M. (1999). DNA sequence similarity requirements for
524 interspecific recombination in *Bacillus*. *Genetics* 153, 1525-1533.
- 525 30. Lovett, S.T., Hurley, R.L., Sutter, V.A., Jr., Aubuchon, R.H., and Lebedeva, M.A.
526 (2002). Crossing over between regions of limited homology in *Escherichia coli*. *RecA*-
527 dependent and *RecA*-independent pathways. *Genetics* 160, 851-859.
- 528 31. Baharoglu, Z., Bikard, D., and Mazel, D. (2010). Conjugative DNA transfer induces
529 the bacterial SOS response and promotes antibiotic resistance development through
530 integron activation. *PLoS Genet* 6, e1001165.
- 531 32. Sorek, R., Zhu, Y., Creevey, C.J., Francino, M.P., Bork, P., and Rubin, E.M. (2007).
532 Genome-wide experimental determination of barriers to horizontal gene transfer.
533 *Science* 318, 1449-1452.
- 534 33. Tuller, T., Girshovich, Y., Sella, Y., Kreimer, A., Freilich, S., Kupiec, M., Gophna, U.,
535 and Ruppin, E. (2011). Association between translation efficiency and horizontal
536 gene transfer within microbial communities. *Nucleic Acids Res* 39, 4743-4755.
- 537 34. Porse, A., Schou, T.S., Munck, C., Ellabaan, M.M.H., and Sommer, M.O.A. (2018).
538 Biochemical mechanisms determine the functional compatibility of heterologous
539 genes. *Nat Commun* 9, 522.
- 540 35. Szabova, J., Ruzicka, P., Verner, Z., Hampl, V., and Lukes, J. (2011). Experimental
541 examination of EFL and MATX eukaryotic horizontal gene transfers: coexistence of
542 mutually exclusive transcripts predates functional rescue. *Mol Biol Evol* 28, 2371-
543 2378.
- 544 36. Pal, C., Papp, B., and Lercher, M.J. (2005). Adaptive evolution of bacterial metabolic
545 networks by horizontal gene transfer. *Nat Genet* 37, 1372-1375.
- 546 37. Rivera, M.C., Jain, R., Moore, J.E., and Lake, J.A. (1998). Genomic evidence for two
547 functionally distinct gene classes. *Proc Natl Acad Sci U S A* 95, 6239-6244.
- 548 38. Cohen, O., Gophna, U., and Pupko, T. (2011). The complexity hypothesis revisited:
549 connectivity rather than function constitutes a barrier to horizontal gene transfer. *Mol*
550 *Biol Evol* 28, 1481-1489.
- 551 39. Jain, R., Rivera, M.C., and Lake, J.A. (1999). Horizontal gene transfer among
552 genomes: the complexity hypothesis. *Proc Natl Acad Sci U S A* 96, 3801-3806.

- 553 40. Baltrus, D.A. (2013). Exploring the costs of horizontal gene transfer. *Trends in Ecology & Evolution* 28, 489-495.
- 554
- 555 41. San Millan, A., Toll-Riera, M., Qi, Q., and MacLean, R.C. (2015). Interactions
556 between horizontally acquired genes create a fitness cost in *Pseudomonas*
557 *aeruginosa*. *Nat Commun* 6, 6845.
- 558 42. Harrison, E., and Brockhurst, M.A. (2012). Plasmid-mediated horizontal gene transfer
559 is a coevolutionary process. *Trends in Microbiology* 20, 262-267.
- 560 43. Brown, C.J., Sen, D., Yano, H., Bauer, M.L., Rogers, L.M., Van der Auwera, G.A.,
561 and Top, E.M. (2013). Diverse broad-host-range plasmids from freshwater carry few
562 accessory genes. *Appl Environ Microbiol* 79, 7684-7695.
- 563 44. Lopatkin, A.J., Meredith, H.R., Srimani, J.K., Pfeiffer, C., Durrett, R., and You, L.
564 (2017). Persistence and reversal of plasmid-mediated antibiotic resistance. *Nat*
565 *Commun* 8, 1689.
- 566 45. Hall, J.P., Wood, A.J., Harrison, E., and Brockhurst, M.A. (2016). Source-sink
567 plasmid transfer dynamics maintain gene mobility in soil bacterial communities. *Proc*
568 *Natl Acad Sci U S A* 113, 8260-8265.
- 569 46. Stevenson, C., Hall, J.P., Harrison, E., Wood, A., and Brockhurst, M.A. (2017). Gene
570 mobility promotes the spread of resistance in bacterial populations. *Isme J* 11, 1930-
571 1932.
- 572 47. Fox, R.E., Zhong, X., Krone, S.M., and Top, E.M. (2008). Spatial structure and
573 nutrients promote invasion of IncP-1 plasmids in bacterial populations. *Isme J* 2,
574 1024-1039.
- 575 48. Bahl, M.I., Hansen, L.H., and Sorensen, S.J. (2007). Impact of conjugal transfer on
576 the stability of IncP-1 plasmid pKJK5 in bacterial populations. *FEMS Microbiol Lett*
577 266, 250-256.
- 578 49. Lopatkin, A.J., Huang, S., Smith, R.P., Srimani, J.K., Sysoeva, T.A., Bewick, S.,
579 Karig, D.K., and You, L. (2016). Antibiotics as a selective driver for conjugation
580 dynamics. *Nat Microbiol* 1, 16044.
- 581 50. Hall, J.P.J., Williams, D., Paterson, S., Harrison, E., and Brockhurst, M.A. (2017).
582 Positive selection inhibits gene mobilisation and transfer in soil bacterial
583 communities. *Nat Ecol Evol* 1, 1348-1353.
- 584 51. Gao, N.L., Zhang, C., Zhang, Z., Hu, S., Lercher, M.J., Zhao, X.M., Bork, P., Liu, Z.,
585 and Chen, W.H. (2018). MVP: a microbe-phage interaction database. *Nucleic Acids*
586 *Res* 46, D700-D707.
- 587 52. Hyman, P., and Abedon, S.T. (2010). Bacteriophage host range and bacterial
588 resistance. *Adv Appl Microbiol* 70, 217-248.
- 589 53. Jain, A., and Srivastava, P. (2013). Broad host range plasmids. *FEMS Microbiol Lett*
590 348, 87-96.
- 591 54. Halary, S., Leigh, J.W., Cheaib, B., Lopez, P., and Baptiste, E. (2010). Network
592 analyses structure genetic diversity in independent genetic worlds. *Proc Natl Acad*
593 *Sci U S A* 107, 127-132.
- 594 55. Sheppard, A.E., Stoesser, N., Wilson, D.J., Sebra, R., Kasarskis, A., Anson, L.W.,
595 Giess, A., Pankhurst, L.J., Vaughan, A., Grim, C.J., et al. (2016). Nested Russian
596 Doll-Like Genetic Mobility Drives Rapid Dissemination of the Carbapenem
597 Resistance Gene blaKPC. *Antimicrob Agents Chemother* 60, 3767-3778.
- 598 56. He, S., Chandler, M., Varani, A.M., Hickman, A.B., Dekker, J.P., and Dyda, F. (2016).
599 Mechanisms of Evolution in High-Consequence Drug Resistance Plasmids. *MBio* 7.
- 600 57. Greated, A., Lambertsen, L., Williams, P.A., and Thomas, C.M. (2002). Complete
601 sequence of the IncP-9 TOL plasmid pWW0 from *Pseudomonas putida*. *Environ*
602 *Microbiol* 4, 856-871.
- 603 58. Nakamura, Y., Itoh, T., Matsuda, H., and Gojobori, T. (2004). Biased biological
604 functions of horizontally transferred genes in prokaryotic genomes. *Nat Genet* 36,
605 760-766.

- 606 59. Porse, A., Schonning, K., Munck, C., and Sommer, M.O.A. (2016). Survival and
607 Evolution of a Large Multidrug Resistance Plasmid in New Clinical Bacterial Hosts.
608 *Mol Biol Evol* 33, 2860-2873.
- 609 60. Silva, J.B., Storms, Z., and Sauvageau, D. (2016). Host receptors for bacteriophage
610 adsorption. *Fems Microbiol Lett* 363.
- 611 61. San Millan, A., Pena-Miller, R., Toll-Riera, M., Halbert, Z.V., McLean, A.R., Cooper,
612 B.S., and MacLean, R.C. (2014). Positive selection and compensatory adaptation
613 interact to stabilize non-transmissible plasmids. *Nat Commun* 5, 5208.
- 614 62. Harrison, E., Dytham, C., Hall, J.P., Guymer, D., Spiers, A.J., Paterson, S., and
615 Brockhurst, M.A. (2016). Rapid compensatory evolution promotes the survival of
616 conjugative plasmids. *Mob Genet Elements* 6, e1179074.
- 617 63. Loftie-Eaton, W., Bashford, K., Quinn, H., Dong, K., Millstein, J., Hunter, S.,
618 Thomason, M.K., Merrih, H., Ponciano, J.M., and Top, E.M. (2017). Compensatory
619 mutations improve general permissiveness to antibiotic resistance plasmids. *Nat Ecol*
620 *Evol* 1, 1354-1363.
- 621 64. Harrison, E., Guymer, D., Spiers, A.J., Paterson, S., and Brockhurst, M.A. (2015).
622 Parallel compensatory evolution stabilizes plasmids across the parasitism-mutualism
623 continuum. *Current biology : CB* 25, 2034-2039.
- 624 65. Yano, H., Wegrzyn, K., Loftie-Eaton, W., Johnson, J., Deckert, G.E., Rogers, L.M.,
625 Konieczny, I., and Top, E.M. (2016). Evolved plasmid-host interactions reduce
626 plasmid interference cost. *Mol Microbiol* 101, 743-756.
- 627 66. Bottery, M.J., Wood, A.J., and Brockhurst, M.A. (2019). Temporal dynamics of
628 bacteria-plasmid coevolution under antibiotic selection. *Isme J* 13, 559-562.
- 629 67. Bottery, M.J., Wood, A.J., and Brockhurst, M.A. (2017). Adaptive modulation of
630 antibiotic resistance through intragenomic coevolution. *Nat Ecol Evol* 1, 1364-1369.
- 631 68. Porse, A., Schonning, K., Munck, C., and Sommer, M.O. (2016). Survival and
632 Evolution of a Large Multidrug Resistance Plasmid in New Clinical Bacterial Hosts.
633 *Mol Biol Evol* 33, 2860-2873.
- 634 69. Turner, P.E., Williams, E.S., Okeke, C., Cooper, V.S., Duffy, S., and Wertz, J.E.
635 (2014). Antibiotic resistance correlates with transmission in plasmid evolution.
636 *Evolution* 68, 3368-3380.
- 637 70. Bobay, L.M., Touchon, M., and Rocha, E.P.C. (2014). Pervasive domestication of
638 defective prophages by bacteria. *P Natl Acad Sci USA* 111, 12127-12132.
- 639 71. Kottara, A., Hall, J.P., Harrison, E., and Brockhurst, M.A. (2016). Multi-host
640 environments select for host-generalist conjugative plasmids. *BMC evolutionary*
641 *biology* 16, 70.
- 642 72. Agrawal, N., Dasaradhi, P.V., Mohammed, A., Malhotra, P., Bhatnagar, R.K., and
643 Mukherjee, S.K. (2003). RNA interference: biology, mechanism, and applications.
644 *Microbiol Mol Biol Rev* 67, 657-685.
- 645 73. Lucchini, S., Rowley, G., Goldberg, M.D., Hurd, D., Harrison, M., and Hinton, J.C.
646 (2006). H-NS mediates the silencing of laterally acquired genes in bacteria. *Plos*
647 *Pathog* 2, e81.
- 648 74. Marraffini, L.A., and Sontheimer, E.J. (2008). CRISPR interference limits horizontal
649 gene transfer in staphylococci by targeting DNA. *Science* 322, 1843-1845.
- 650 75. Dupuis, M.E., Villion, M., Magadan, A.H., and Moineau, S. (2013). CRISPR-Cas and
651 restriction-modification systems are compatible and increase phage resistance. *Nat*
652 *Commun* 4, 2087.
- 653 76. Oliveira, P.H., Touchon, M., and Rocha, E.P. (2016). Regulation of genetic flux
654 between bacteria by restriction-modification systems. *Proc Natl Acad Sci U S A* 113,
655 5658-5663.
- 656 77. Palmer, K.L., and Gilmore, M.S. (2010). Multidrug-resistant enterococci lack
657 CRISPR-cas. *MBio* 1.
- 658 78. Watson, B.N.J., Staals, R.H.J., and Fineran, P.C. (2018). CRISPR-Cas-Mediated
659 Phage Resistance Enhances Horizontal Gene Transfer by Transduction. *MBio* 9.

- 660 79. Goldberg, G.W., Jiang, W., Bikard, D., and Marraffini, L.A. (2014). Conditional
661 tolerance of temperate phages via transcription-dependent CRISPR-Cas targeting.
662 *Nature* 514, 633-637.
- 663 80. Faure, G., Makarova, K.S., and Koonin, E.V. (2019). CRISPR-Cas: Complex
664 Functional Networks and Multiple Roles beyond Adaptive Immunity. *J Mol Biol* 431,
665 3-20.
- 666 81. Gophna, U., Kristensen, D.M., Wolf, Y.I., Popa, O., Drevet, C., and Koonin, E.V.
667 (2015). No evidence of inhibition of horizontal gene transfer by CRISPR-Cas on
668 evolutionary timescales. *Isme J* 9, 2021-2027.
- 669 82. Gao, N.L., Chen, J., Lercher, M.J., and Chen, W.-H. (2018). Prokaryotic genome
670 expansion is facilitated by phages and plasmids but impaired by CRISPR. *BioRxiv*.
- 671 83. Doron, S., Melamed, S., Ofir, G., Leavitt, A., Lopatina, A., Keren, M., Amitai, G., and
672 Sorek, R. (2018). Systematic discovery of antiphage defense systems in the
673 microbial pangenome. *Science* 359.
- 674 84. Thomas, C.M., and Nielsen, K.M. (2005). Mechanisms of, and barriers to, horizontal
675 gene transfer between bacteria. *Nat Rev Microbiol* 3, 711-721.
- 676 85. Berngruber, T.W., Weissing, F.J., and Gandon, S. (2010). Inhibition of superinfection
677 and the evolution of viral latency. *J Virol* 84, 10200-10208.
- 678 86. Faure, G., Shmakov, S.A., Yan, W.X., Cheng, D.R., Scott, D.A., Peters, J.E.,
679 Makarova, K.S., and Koonin, E.V. (2019). CRISPR-Cas in mobile genetic elements:
680 counter-defence and beyond. *Nat Rev Microbiol*.
- 681 87. Hartl, D.L., and Clark, A.G. (2007). Principles of population genetics, 4th Edition,
682 (Sunderland, Mass.: Sinauer Associates).
- 683 88. Charlesworth, B. (2009). Effective population size and patterns of molecular evolution
684 and variation. *Nat Rev Genet* 10, 195-205.
- 685 89. San Millan, A., and MacLean, R.C. (2017). Fitness Costs of Plasmids: a Limit to
686 Plasmid Transmission. *Microbiol Spectr* 5.
- 687 90. Vogwill, T., and MacLean, R.C. (2015). The genetic basis of the fitness costs of
688 antimicrobial resistance: a meta-analysis approach. *Evol Appl* 8, 284-295.
- 689 91. Price, M.N., Wetmore, K.M., Waters, R.J., Callaghan, M., Ray, J., Liu, H., Kuehl, J.V.,
690 Melnyk, R.A., Lamson, J.S., Suh, Y., et al. (2018). Mutant phenotypes for thousands
691 of bacterial genes of unknown function. *Nature* 557, 503-509.
- 692 92. van Opijnen, T., and Camilli, A. (2013). Transposon insertion sequencing: a new tool
693 for systems-level analysis of microorganisms. *Nat Rev Microbiol* 11, 435-442.
- 694 93. Giaever, G., Chu, A.M., Ni, L., Connelly, C., Riles, L., Veronneau, S., Dow, S.,
695 Lucau-Danila, A., Anderson, K., Andre, B., et al. (2002). Functional profiling of the
696 *Saccharomyces cerevisiae* genome. *Nature* 418, 387-391.
- 697 94. Sela, I., Wolf, Y.I., and Koonin, E.V. (2016). Theory of prokaryotic genome evolution.
698 *P Natl Acad Sci USA* 113, 11399-11407.
- 699 95. Bobay, L.M., and Ochman, H. (2018). Factors driving effective population size and
700 pan-genome evolution in bacteria. *Bmc Evol Biol* 18.
- 701 96. Mira, A., Ochman, H., and Moran, N.A. (2001). Deletional bias and the evolution of
702 bacterial genomes. *Trends Genet* 17, 589-596.
- 703 97. Gil, R., Sabater-Munoz, B., Latorre, A., Silva, F.J., and Moya, A. (2002). Extreme
704 genome reduction in *Buchnera* spp.: toward the minimal genome needed for
705 symbiotic life. *Proc Natl Acad Sci U S A* 99, 4454-4458.
- 706 98. Veyrier, F.J., Dufort, A., and Behr, M.A. (2011). The rise and fall of the
707 *Mycobacterium tuberculosis* genome. *Trends Microbiol* 19, 156-161.
- 708 99. McInerney, J.O., McNally, A., and O'Connell, M.J. (2017). Why prokaryotes have
709 pangenomes. *Nat Microbiol* 2.
- 710 100. Niehus, R., Mitri, S., Fletcher, A.G., and Foster, K.R. (2015). Migration and horizontal
711 gene transfer divide microbial genomes into multiple niches. *Nature Communications*
712 6.
- 713 101. Andreani, N.A., Hesse, E., and Vos, M. (2017). Prokaryote genome fluidity is
714 dependent on effective population size. *Isme J* 11, 1719-1721.

- 715 102. Collins, R.E., and Higgs, P.G. (2012). Testing the Infinitely Many Genes Model for
716 the Evolution of the Bacterial Core Genome and Pangenome. *Mol Biol Evol* 29, 3413-
717 3425.
- 718 103. Baumdicker, F., Hess, W.R., and Pfaffelhuber, P. (2012). The Infinitely Many Genes
719 Model for the Distributed Genome of Bacteria. *Genome Biol Evol* 4, 443-456.
- 720 104. Rocha, E.P.C. (2018). Neutral Theory, Microbial Practice: Challenges in Bacterial
721 Population Genetics. *Mol Biol Evol* 35, 1338-1347.
- 722 105. May, R.M., and Anderson, R.M. (1983). Epidemiology and Genetics in the
723 Coevolution of Parasites and Hosts. *P Roy Soc Lond a Mat* 390, 219-219.
- 724 106. Partridge, S.R., Kwong, S.M., Firth, N., and Jensen, S.O. (2018). Mobile Genetic
725 Elements Associated with Antimicrobial Resistance. *Clin Microbiol Rev* 31.
- 726 107. Rozwandowicz, M., Brouwer, M.S.M., Fischer, J., Wagenaar, J.A., Gonzalez-Zorn,
727 B., Guerra, B., Mevius, D.J., and Hordijk, J. (2018). Plasmids carrying antimicrobial
728 resistance genes in Enterobacteriaceae. *J Antimicrob Chemother* 73, 1121-1137.
- 729 108. Gardner, A. (2009). Adaptation as organism design. *Biology letters* 5, 861-864.
- 730 109. Karkman, A., Parnanen, K., and Larsson, D.G.J. (2019). Fecal pollution can explain
731 antibiotic resistance gene abundances in anthropogenically impacted environments.
732 *Nat Commun* 10, 80.
- 733 110. Klumper, U., Riber, L., Dechesne, A., Sannazzarro, A., Hansen, L.H., Sorensen, S.J.,
734 and Smets, B.F. (2015). Broad host range plasmids can invade an unexpectedly
735 diverse fraction of a soil bacterial community. *Isme J* 9, 934-945.
- 736 111. Cordero, O.X., and Polz, M.F. (2014). Explaining microbial genomic diversity in light
737 of evolutionary ecology. *Nat Rev Microbiol* 12, 263-273.
- 738 112. Datta, M.S., Sliwerska, E., Gore, J., Polz, M.F., and Cordero, O.X. (2016). Microbial
739 interactions lead to rapid micro-scale successions on model marine particles. *Nat*
740 *Commun* 7, 11965.
- 741 113. Leibold, M.A. (2018). *Metacommunity ecology*, (Princeton, NJ: Princeton University
742 Press).
- 743 114. Makarova, K.S., Wolf, Y.I., Mekhedov, S.L., Mirkin, B.G., and Koonin, E.V. (2005).
744 Ancestral paralogs and pseudoparalogs and their role in the emergence of the
745 eukaryotic cell. *Nucleic Acids Res* 33, 4626-4638.
- 746 115. Doolittle, R.F. (1995). The multiplicity of domains in proteins. *Annu Rev Biochem* 64,
747 287-314.
- 748 116. Nevoigt, E., Fassbender, A., and Stahl, U. (2000). Cells of the yeast *Saccharomyces*
749 *cerevisiae* are transformable by DNA under non-artificial conditions. *Yeast* 16, 1107-
750 1110.
- 751 117. Heinemann, J.A., and Sprague, G.F., Jr. (1989). Bacterial conjugative plasmids
752 mobilize DNA transfer between bacteria and yeast. *Nature* 340, 205-209.
- 753 118. Soanes, D., and Richards, T.A. (2014). Horizontal gene transfer in eukaryotic plant
754 pathogens. *Annu Rev Phytopathol* 52, 583-614.
- 755 119. Monier, A., Pagarete, A., de Vargas, C., Allen, M.J., Read, B., Claverie, J.M., and
756 Ogata, H. (2009). Horizontal gene transfer of an entire metabolic pathway between a
757 eukaryotic alga and its DNA virus. *Genome Res* 19, 1441-1449.
- 758 120. Gallot-Lavallee, L., and Blanc, G. (2017). A Glimpse of Nucleo-Cytoplasmic Large
759 DNA Virus Biodiversity through the Eukaryotic Genomics Window. *Viruses* 9.
- 760 121. Doolittle, W.F. (1998). You are what you eat: a gene transfer ratchet could account
761 for bacterial genes in eukaryotic nuclear genomes. *Trends Genet* 14, 307-311.
- 762 122. Glass, N.L., Jacobson, D.J., and Shiu, P.K. (2000). The genetics of hyphal fusion and
763 vegetative incompatibility in filamentous ascomycete fungi. *Annu Rev Genet* 34, 165-
764 186.
- 765 123. McCarthy, C.G.P., and Fitzpatrick, D.A. (2019). Pan-genome analyses of model
766 fungal species. *Microb Genom* 5.
- 767 124. Read, B.A., Kegel, J., Klute, M.J., Kuo, A., Lefebvre, S.C., Maumus, F., Mayer, C.,
768 Miller, J., Monier, A., Salamov, A., et al. (2013). Pan genome of the phytoplankton
769 *Emiliana underpins its global distribution*. *Nature* 499, 209-213.

- 770 125. Temporini, E.D., and VanEtten, H.D. (2004). An analysis of the phylogenetic
771 distribution of the pea pathogenicity genes of *Nectria haematococca* MPVI supports
772 the hypothesis of their origin by horizontal transfer and uncovers a potentially new
773 pathogen of garden pea: *Neocosmospora boniensis*. *Curr Genet* 46, 29-36.
- 774 126. Coleman, J.J., Rounsley, S.D., Rodriguez-Carres, M., Kuo, A., Wasmann, C.C.,
775 Grimwood, J., Schmutz, J., Taga, M., White, G.J., Zhou, S., et al. (2009). The
776 genome of *Nectria haematococca*: contribution of supernumerary chromosomes to
777 gene expansion. *PLoS Genet* 5, e1000618.
- 778 127. He, C., Rusu, A.G., Poplawski, A.M., Irwin, J.A., and Manners, J.M. (1998). Transfer
779 of a supernumerary chromosome between vegetatively incompatible biotypes of the
780 fungus *Colletotrichum gloeosporioides*. *Genetics* 150, 1459-1466.
- 781