

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/315869060>

Why Prokaryotes Have Pangenomes

Article *in* Nature Microbiology · March 2017

DOI: 10.1038/nmicrobiol.2017.40

CITATIONS

2

READS

88

3 authors, including:



[James McInerney](#)

The University of Manchester

124 PUBLICATIONS 4,015 CITATIONS

[SEE PROFILE](#)



[Mary J O'Connell](#)

University of Leeds

96 PUBLICATIONS 883 CITATIONS

[SEE PROFILE](#)

Some of the authors of this publication are also working on these related projects:



Horizontal gene transfer [View project](#)



A hologenomic and evolutionary approach to determine large-scale molecular adaptations associated with toxin sequestration from diet in Poison Arrow Frogs. [View project](#)

All content following this page was uploaded by [James McInerney](#) on 21 April 2017.

The user has requested enhancement of the downloaded file.

Why prokaryotes have pangenomes

James O. McInerney^{1*}, Alan McNally² and Mary J. O'Connell³

The existence of large amounts of within-species genome content variability is puzzling. Population genetics tells us that fitness effects of new variants—either deleterious, neutral or advantageous—combined with the long-term effective population size of the species determines the likelihood of a new variant being removed, spreading to fixation or remaining polymorphic. Consequently, we expect that selection and drift will reduce genetic variation, which makes large amounts of gene content variation in some species so puzzling. Here, we amalgamate population genetic theory with models of horizontal gene transfer and assert that pangenomes most easily arise in organisms with large long-term effective population sizes, as a consequence of acquiring advantageous genes, and that the focal species has the ability to migrate to new niches. Therefore, we suggest that pangenomes are the result of adaptive, not neutral, evolution.

As soon as different strains of the same species had their genomes sequenced, it became apparent that there was enormous intraspecific variability in prokaryotic genome content¹. Indeed, terms such as 'core' and 'accessory' genomes have been coined in order to describe this variation². The core genome refers to 'essential' gene families that are found in all members sequenced thus far, and the accessory genome refers to 'dispensable' genes that are not in every genome³. The pangenome consists of all the gene families that have been found in the species as a whole⁴ (see Fig. 1; for full definitions of these terms and others used in this Perspective, see Box 1). Some prokaryotic species have extensive (or open) pangenomes, while others have genomes that manifest very few gene content differences (closed pangenomes). Our understanding of the pangenome of a species will depend on whether we have sampled the broad diversity of the species and how many genomes we have sequenced from this diversity. The dominant source of genome content variability for prokaryotes is horizontal gene transfer (HGT), allied to differential gene losses, with gene duplications also playing a role, albeit a lesser one⁵. However, the absence of theory to explain pangenomes is a gap in the new synthesis. In this Perspective, we present testable theory governing pangenome accumulation and present our predictions for future empirical observations.

Non-treelike evolution of genomes

Nearly three decades ago, Martinez-Murcia *et al.*⁶ observed incongruence between near-identical 16S rRNA gene sequences in the genus *Aeromonas* and low levels of DNA:DNA hybridization. Though unusual, this disparity was not attributed to the idea of a pangenome, since the genome sequences were unknown at that time. Soon, however, it became clear that prokaryotic genomes were substantially affected by HGT^{7,8}, calling into question the previously unshakeable tree of life hypothesis, though some still felt that HGT did not affect phylogenies⁹.

Today, the thousands of prokaryotic genome sequences available reveal the pervasive influence of introgressions of many kinds¹⁰. The largest pangenome analysis for a single species to date included 2,085 *Escherichia coli* genomes¹¹, and this estimated 3,188 core gene families (which they defined as being present in 95% of genomes)

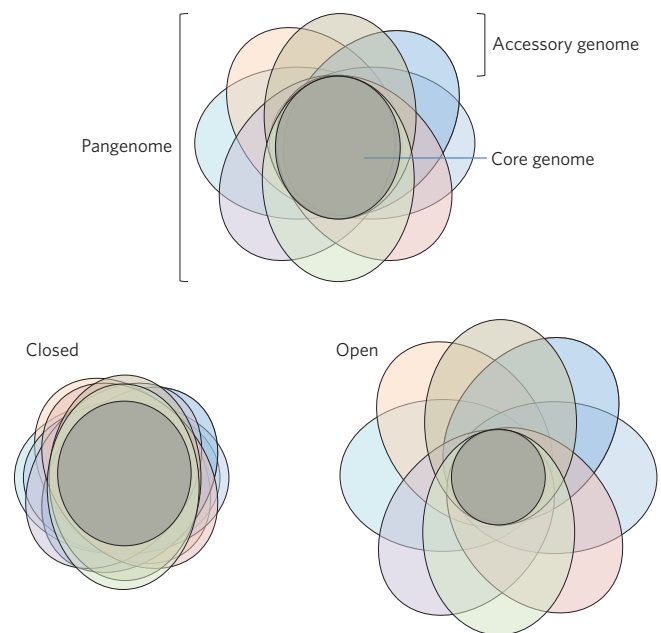


Figure 1 | Schematic representation of pangenomes as Venn diagrams.

Species differ in the sizes of their pangenomes, with larger, more open pangenomes correlating with larger long-term effective population sizes and the ability to migrate.

and approximately 90,000 unique gene families. By contrast, the intracellular pathogen *Chlamydia trachomatis* has a pangenome size only slightly bigger than its core genome (974 gene pangenome, 821 gene core genome), with 67 genomes sequenced (see Table 1). This gives us a range of core genome size, from 3% to 84% for well-sampled genomes. As more genomes are collected, the core genome tends to get smaller and the accessory genome tends to get bigger¹², and continued sequencing will change these numbers. Interestingly, exploring the pattern of gene presence and absence in a sample of 573 genomes and then extrapolating to a larger number

¹Division of Evolution and Genomic Sciences, School of Biological Sciences, Faculty of Biology, Medicine and Health, University of Manchester, Manchester Academic Health Science Centre, Manchester M13 9PL, UK. ²Institute of Microbiology and Infection, College of Medical and Dental Sciences, University of Birmingham, Edgbaston, Birmingham B15 2TT, UK. ³Computational and Molecular Evolutionary Biology Group, School of Biology, Faculty of Biological Sciences, University of Leeds, Leeds LS2 9JT, UK. *e-mail: james.mcinerney@manchester.ac.uk

Box 1 | Glossary.

Core genome. The collection of gene families that are found to be present in all members of a particular species.

Accessory genome: the collection of gene families that are found in some, but not all, genomes of a particular species.

Pangenome. The entire collection of gene families that are found in a given species.

Exogenous DNA. DNA that can be found outside cells. This is usually DNA from dead cells or mobile genetic elements.

Horizontal gene transfer: the transfer of a gene from one organism to another organism, where the recipient is not a direct descendent of the donor.

New synthesis. Refers to the reconciliation of Darwinian evolution with the Mendelian laws of heredity. Also known as the modern evolutionary synthesis, it consists of a conceptual framework underpinned by mathematics and empirical observations that explain the evolution of life on the planet.

16S rRNA. The RNA molecule that is found in the small subunit of the ribosome. The gene encoding this RNA molecule has been used extensively for phylogenetic analysis.

Tree of life hypothesis. The hypothesis that all cellular life on the planet can be depicted on a single phylogenetic tree. The alternative hypothesis is that living systems frequently exchange genes and life is poorly described by a tree, but better described as a network.

Random genetic drift. Genetic drift refers to changes in gene

frequency from one generation to the next due to the random sampling of individuals that successfully reproduce.

Neutral substitution. Neutral genetic changes are those changes that have no effect on the fitness of an organism. Natural selection does not act on these variants.

Nearly neutral. Nearly neutral alleles do confer a fitness difference on the individual with the new variant, however, this difference is not sufficient to overcome genetic drift. In this case, though there is a fitness difference, fixation of the new variant is still determined by drift, not selection.

Addictive genetic elements. Genes that result in cell death if they are lost during cellular replication. The classic example is a toxin–antitoxin system, where a long-lived toxin and a short-lived antitoxin exist together. Losing either or both genes results in depletion of the antitoxin and the toxin killing the host cell. This means the cells are ‘addicted’ to the system.

Selfish genetic elements. Parasitic genes or collections of genes whose primary objective is to replicate while providing little, if any, benefit to their hosts.

Selective sweep. The situation when a new variant gene or genome arises that results in an increase in fitness of the carrier, causing a rapid rise to fixation in the population. This results in a reduction in genetic variation near the new mutation or even in the species as a whole.

of genomes, the entire bacterial pangenome has been estimated to be infinite in size¹³. This has been likened to a “constant rain of genetic material on genomes”¹³, and implies that genomes have an almost limitless supply of genes from which they can sample.

Pangenomes can also be found in eukaryotes (Table 1). For example, the human pangenome is thought to have between 15–40 Mb of accessory DNA, approximately 0.5–1.3%¹⁴, while the 14 genomes of the coccolithophore *Emiliania huxleyi* have only 69.5% of identified genes common to all genomes. However, in eukaryotes, gene inheritance is somewhat different, with lower HGT levels than in prokaryotes¹⁵ and higher levels of gene duplication¹⁶. In this Perspective, we focus on prokaryotes in part because we have not yet sampled as much intraspecific genome-level variation across a broad range of eukaryotes as we have for prokaryotes.

The processes leading to the generation of pangenomes still requires a thorough theoretical explanation, and one that incorporates the fact that there is a distribution of pangenome sizes, from minimal to extensive. HGT is a form of mutation and can be treated as such in models of pangenome evolution. These models also have to take into account variation in effective population size (N_e , defined as the number of individuals that contribute offspring to the next generation), mutation rates, selection coefficients, influence of random drift, and kinds of speciation, and there is also variation in the tendency of a particular prokaryotic species to form extensive pangenomes (Table 1). Delivering new alleles or genes into a cell (the baseline rate of introgression) is not sufficient to ensure their retention (the realised rate)¹⁷. We can assume, from the plentitude of mobile genetic elements and exogenous DNA, that gene delivery is quite frequent; the question is what promotes retention and why there is not a typical genome for every prokaryotic species.

Our model for how prokaryotic pangenomes arise and are maintained is based on the existence of widespread and numerous cryptic niches combined with natural selection for beneficial genome types (see Box 2 for a note on small selective pressures). This model has a growing amount of support from empirical data. We also explain why other models provide inadequate theory for pangenomes.

Random drift model

Firstly, we consider a model where drift is not overcome by natural selection and where newly acquired genes are neutral or nearly neutral. Evolutionary theory tells us that the fate of a new allele in a population is dependent on the long-term effective population size of the species and the fitness effect of the new allele¹⁸. A truly neutral new allele in a population of size N will have an initial frequency of $1/N$. If the underlying acquisition rate of new alleles is μ , then the rate of fixation of new alleles purely by drift is $N\mu \times 1/N = \mu$. This means that the probability of fixation of neutral newly acquired alleles is independent of population size and equal to the rate of introduction of the alleles¹⁹. The time to fixation of neutral alleles is, on average, equal to $2N$, meaning that a neutral allele could remain polymorphic and at low frequency in a large population for a long time. Therefore, this model could potentially explain the existence of extensive pangenomes. However, unlike a single point mutation that simply changes the identity of the encoded nucleotide, a new protein-coding gene, say, 1,000 nucleotides long, requires a certain amount of energy in order to be replicated, transcribed and translated²⁰. We expect few transferred genetic segments to achieve the perfect balance of functional benefit, offsetting the cost of production and maintenance of this function. For a non-neutral allele with a selective coefficient (s) to be fixed in the population by drift, it must satisfy the condition that $|s| \ll 1/N$, that is, a nearly neutral allele, *sensu* Ohta²¹. For organisms with large N_e , s would have to be very close to zero in order to ever become fixed or indeed to remain polymorphic for a long period of time. Additionally, if the processes of acquisition and maintenance were truly neutral for the majority of genes, then some genomes might expand and become as large as eukaryotic genomes, but instead, prokaryotic genomes generally remain in the range of 1–8 Mb (ref. 22). Indeed, empirical genome analyses have demonstrated that prokaryotic genomes are biased towards deletion of DNA²³, indicating that this bias would tend to delete neutral alleles and again we would not see pangenomes. Clearly, a neutral model for pangenome accrual will not work. In any case, recent simulation work has shown that, on average, HGTs in prokaryotes tend to be adaptive²⁴.

Table 1 | A short list of pangenome, core genome and accessory genome sizes for a variety of prokaryotes and eukaryotes.

Name	No.	Core genome (% of total)	Accessory genome (% of total)	Pangenome size	No. of new genes per genome* (% of core)	Lifestyle
<i>Chlamydia trachomatis</i>	67	821 (84)	153 (15)	974	2 (0.2)	Obligate intracellular
<i>Bacillus anthracis</i>	42	4,001 (65)	2,065 (35)	6,066	49 (1.2)	Obligate pathogen
<i>Corynebacterium pseudotuberculosis</i>	40	1,205 (53)	1,041 (46)	2,246	26 (2.5)	Facultative intracellular pathogen
<i>Streptococcus pyogenes</i>	49	986 (37)	1,644 (62)	2,683	34 (3.4)	Human-restricted pathogen
<i>Staphylococcus aureus</i>	74	1,639 (36)	2,879 (64)	4,520	39 (2.3)	Commensal with humans and other animals
<i>Corynebacterium diphtheriae</i>	13	1,632 (34)	3,156 (66)	4,786	243 (14.9)	Free-living
<i>Helicobacter pylori</i>	83	708 (28)	1,770 (71)	2,478	21 (3)	Intercellular, mucus-associated pathogen
<i>Acinetobacter baumannii</i>	40	1,997 (28)	5,042 (71)	7,039	126 (6.3)	Soil-dweller; opportunistic pathogen
<i>Bacillus cereus</i>	30	3,245 (27)	8,666 (72)	11,911	288 (8.9)	Associated with soil, rumen, guts of mammals and insects; occasionally pathogenic
<i>Klebsiella pneumoniae</i>	63	3,062 (26)	8,456 (74)	11,518	134 (4.4)	Free-living, azotroph, pathogen
<i>Prochlorococcus marinus</i>	40	1,175 (18)	5,260 (81)	6,435	132 (11.2)	Photosynthetic picoplankton
<i>Salmonella enterica</i>	206	1,717 (16)	8,550 (83)	10,267	42 (2.4)	Associated with rumen, water, soil; pathogen
<i>Oryza sativa</i> ³⁸	3	37,200 (92)	3,162 (8)	40,362	1,054 (2.8)	Semi-aquatic, agricultural
<i>Glycine soja</i> ³⁹	7	28,716 (80.1)	7,134 (19.9)	35,850	1,019 (3.5)	Agricultural
<i>Emiliania huxleyi</i> ⁴⁰	14	20,084 (69.5)	8,811 (30.5)	28,895	629 (3)	Aquatic, global distribution

Prokaryote genome data was obtained from the pangenome.de website⁴¹ (<http://pangenome.tuebingen.mpg.de/>). Eukaryotic data was calculated directly from the papers cited. *Average values.

Box 2 | Small selective pressures.

Selection for mutations that confer even very small fitness effects can be seen in organisms with large N_e (ref. 29). *E. coli* has a large pangenome and N_e in this species is estimated to be 25,000,000 (ref. 29). As a consequence, very weak selective effects can overcome genetic drift in *E. coli*. The best-known example can be seen in the way in which translational selection for codon usage in highly expressed genes matches with the cellular abundances of cognate transfer RNAs⁴². For instance, in highly expressed genes, *E. coli* uses the phenylalanine UUC codon more than twice as often as the UUU codon, demonstrating that this very weak selective pressure is capable of overcoming genetic drift in *E. coli*, though only in highly expressed genes⁴². *E. coli*, with its large N_e , is very sensitive to small selective differences caused by mutations in its genes. This includes transcriptional and translational selection⁴³, as well as selection for function³¹ and the cost of maintenance²⁰. Not all organisms have a very large N_e , however. Particular pathogens or symbionts that frequently encounter bottlenecks during transmission have a small N_e (ref. 29). The obligate intracellular pathogen *Mycoplasma genitalium*, which probably has a small N_e , does not show evidence of translational selection in any genes, whether expressed at high or low levels⁴⁴. In the human genome, drift is not overcome by selection for codon usage optimisation⁴⁵. To put it another way, N_e plays a key role in determining whether selective pressures are able to influence evolutionary outcomes, with the genomes of organisms with large populations showing extreme sensitivity to even the smallest selective differences.

Models with associated fitness costs

Another potential explanation for pangenomes is that accessory genes are composed largely of selfish or addictive genetic elements, and extreme genome variability exists because genomes cannot get rid of these selfish elements, even if they are deleterious. However, analysis of the functions of the accessory genomes do not provide support for this scenario²⁵. Of course, some accessory genes are selfish elements, such as phage or toxin-antitoxin genes²⁶, but thousands of known accessory genes have other known functions²² and do not appear to have 'addictive' traits, so a theory based on selfish genes is insufficient here (see Fig. 2 for accessory gene analysis of 228 *E. coli* ST131 genomes)²⁷.

Comparison of closely related genomes indicates that many HGTs are relatively transient, being frequently supplanted by other newcomers¹². This might suggest that new genes are typically deleterious. Baltrus has explored the costs of HGT, including the disruption of genomes, the cytotoxic effects of HGT, the energetic cost of having additional DNA as well as its transcription and translation, the potential for HGT to disrupt various intracellular interactions, and the system-level effects of having additional protein products in a cell²⁸. However, while HGT can have these costs, if HGT was always deleterious, or even usually deleterious, it could not result in pangenomes. Additionally, the knock-on effect would be to promote the evolution of lower HGT rates¹⁸. It is clear that HGT rates, at least in some organisms, are quite high²⁴, suggesting that HGT is not always deleterious.

An adaptive model

We suggest that HGT genes are largely—though not always—adaptive, and the presence of pangenomes is typically an adaptive phenomenon though not in the sense of selective sweeps. Standard

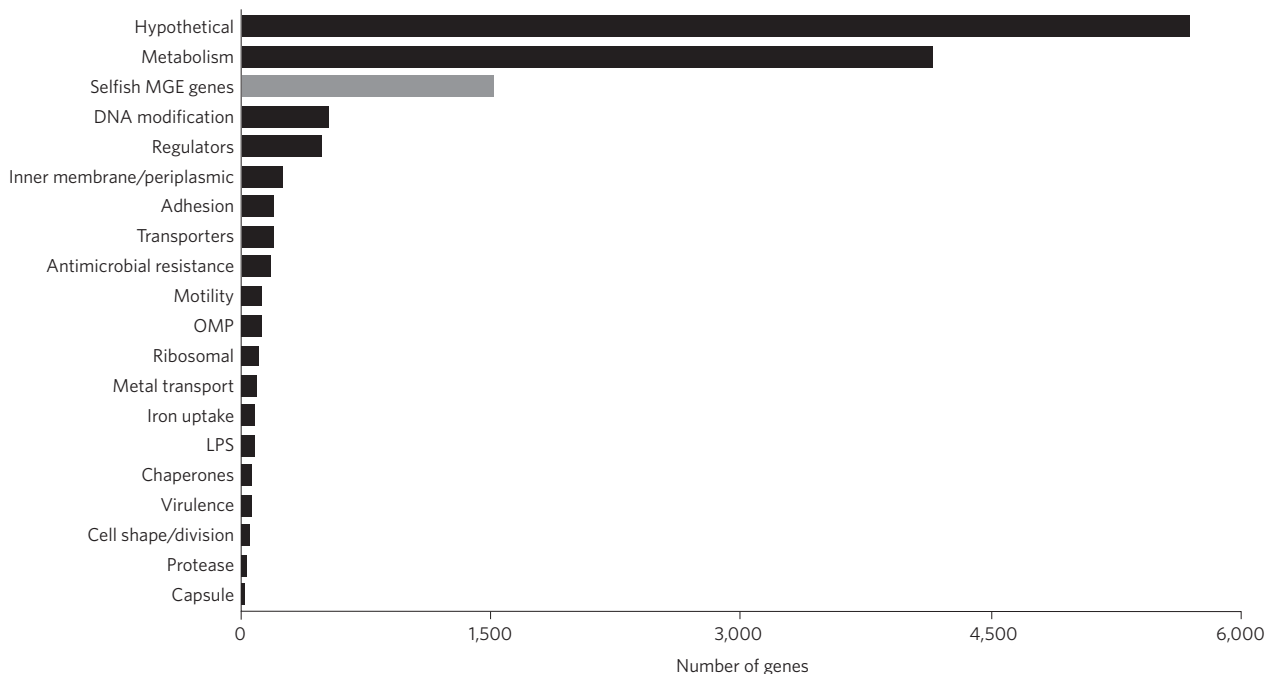


Figure 2 | Analysis of accessory gene functions in 228 *E. coli* ST131 genomes. Though selfish elements constitute a large portion of the known functions, they are not the majority. MGE, mobile genetic elements; OMP, outer membrane protein; LPS, lipopolysaccharide.

evolutionary theory states that the introduction of a new advantageous allele and its fixation by natural selection (a selective sweep) tends to reduce variability in a population, even in the presence of recombination²⁹. So, at first glance, an adaptive model would seem an unlikely explanation for pangenomes. The problem lies with the simplicity of that particular model.

A new compartment model by Niehus *et al.*³⁰ that explicitly models HGT and migration has shown the plausibility of selection on HGT genes driving population differentiation. Using a mathematical approach, the authors showed that in the case of a selectively advantageous HGT event, diversity is removed from the species when there is no migration into or out of the compartment or niche occupied by the focal prokaryotic community. By contrast, a model that includes migration to and from the niche, combined with HGT of a selectively advantageous gene, can theoretically result in a situation where diversity is not necessarily reduced. While this model does not specifically deal with the issue of pangenomes, it does show that diversity within a species can be maintained if advantageous HGT occurs, provided migration can also occur in that species³⁰. Migration might be easy for species such as *E. coli*, that can move, say, from one gastrointestinal tract to another, but perhaps less so for species such as *C. trachomatis*, an intracellular parasite for which new variants must compete *in situ* with wild types. In addition, as the earlier discussion on codon usage showed (see Box 2), selection overcoming drift in prokaryotes is crucially dependent on N_e for the species.

For the Niehus *et al.* model to work, we would need empirical evidence that ostensibly dispensable genes are commonly advantageous. There is a growing body of evidence that accessory genes might provide significant benefit^{31,32}. Karcagi and co-workers analysed a range of *E. coli* genomes at different levels of gene deletion, specifically genes that had been recently acquired by HGT³¹. They found that HGT genes conferred significant benefits in terms of substrate utilisation, efficiency of resource usage to build new cells, and tolerance to stress. Loss of HGT genes tended to affect fitness in several measurable ways, including the induction of a general stress response, inability to grow at all in

some environmental conditions, reduction in growth rate in others, and loss of efficiency of substrate utilisation. The authors concluded that any advantage of DNA loss in terms of a reduction in the cost of replication, transcription and translation was minimal, and was generally overcome by the disadvantage of losing the actual sequences and their encoded functions. Hutchison *et al.*³² constructed a minimal prokaryote genome and demonstrated that significant numbers of genes of unknown function are absolutely essential for life in their minimal genome. Though these essential genes are not universal across life, it is likely that extensive epistatic interactions and dependencies will exist for any system and context-dependent gene loss is frequently deleterious. What this minimal genome shows is that seemingly dispensable genes are not always dispensable, and also that there is still a lot we do not know about gene dependencies.

With this model, we do not suggest that selection can only favour gene gain. Though prokaryotic genomes can grow in size to overlap eukaryotic genome sizes³³, gene loss is obviously just as important as gain, and genes that are not relevant for the ecological niche in which an organism finds itself will soon be lost. Lee and Marx³⁴ have shown selection-driven genome reduction in *Methylobacterium extorquens* AM1 experimental populations. Further investigation revealed a decreased performance of reduced-genome *M. extorquens* AM1 outside the environment in which the deletions were selected, indicating again that accessory genomes can be hugely beneficial, but that context and niche are important. In one environment, deletions are advantageous for a species, but in another, acquisitions provide the advantage.

Conclusion

In conclusion, we infer that effective population size and the ability to migrate to new niches are the most influential factors in determining pangenome size. From Table 1, we can see a strong correlation between lifestyle and the percentage of genes in the core genome of a species. At one extreme, the obligate intracellular pathogen *C. trachomatis* has a core genome of 84% of its pangenome, while the prokaryote thought to be the most abundant on

the planet, *Prochlorococcus marinus*, has a core genome of only 18% of its pangenome, and with each new genome of *P. marinus* sequenced, the new gene discovery rate is at 11.2% of the core genome size. An additional corollary of selection–migration driven pangenomes is that the number of ecological niches on the planet must be enormous. Recent analysis of genomic diversity has suggested that there are one trillion (10^{12}) microbial species on Earth³⁵, which implies the existence of a similar number of ecological niches.

That the majority of genes in the biosphere are not strongly attached to any group of organisms has been a surprise of the genomics era, and consequently this ‘public goods’ hypothesis needed explanation^{36,37}. Future empirical work will involve understanding the precise interplay between HGT, selection, drift, migration, population size and pangenomes.

Received 2 August 2016; accepted 22 February 2017;
published 28 March 2017

References

- Perna, N. T. *et al.* Genome sequence of enterohaemorrhagic *Escherichia coli* O157:H7. *Nature* **409**, 529–533 (2001).
- Young, J. P. *et al.* The genome of *Rhizobium leguminosarum* has recognizable core and accessory components. *Genome Biol.* **7**, R34 (2006).
- Tettelin, H. *et al.* Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial “pan-genome”. *Proc. Natl Acad. Sci. USA* **102**, 13950–13955 (2005).
- Ku, C. *et al.* Endosymbiotic gene transfer from prokaryotic pangenomes: inherited chimerism in eukaryotes. *Proc. Natl Acad. Sci. USA* **112**, 10139–10146 (2015).
- Treangen, T. J. & Rocha, E. P. Horizontal transfer, not duplication, drives the expansion of protein families in prokaryotes. *PLoS Genet.* **7**, e1001284 (2011).
- Martinez-Murcia, A. J., Benlloch, S. & Collins, M. D. Phylogenetic interrelationships of members of the genera *Aeromonas* and *Plesiomonas* as determined by 16S ribosomal DNA sequencing: lack of congruence with results of DNA–DNA hybridizations. *Int. J. Syst. Bacteriol.* **42**, 412–421 (1992).
- Creevey, C. J. *et al.* Does a tree-like phylogeny only exist at the tips in the prokaryotes? *Proc. R. Soc. Lond. B* **271**, 2551–2558 (2004).
- Doolittle, W. F. Phylogenetic classification and the universal tree. *Science* **284**, 2124–2129 (1999).
- Daubin, V., Moran, N. A. & Ochman, H. Phylogenetics and the cohesion of bacterial genomes. *Science* **301**, 829–832 (2003).
- Bapteste, E. *et al.* Evolutionary analyses of non-genealogical bonds produced by introgressive descent. *Proc. Natl Acad. Sci. USA* **109**, 18266–18272 (2012).
- Land, M. *et al.* Insights from 20 years of bacterial genome sequencing. *Funct. Integr. Genomics* **15**, 141–161 (2015).
- Lukjancenko, O., Wassenaar, T. M. & Ussery, D. W. Comparison of 61 sequenced *Escherichia coli* genomes. *Microb. Ecol.* **60**, 708–720 (2010).
- Lapierre, P. & Gogarten, J. P. Estimating the size of the bacterial pan-genome. *Trends Genet.* **25**, 107–110 (2009).
- Li, R. *et al.* Building the sequence map of the human pan-genome. *Nat. Biotechnol.* **28**, 57–63 (2010).
- Ku, C. *et al.* Endosymbiotic origin and differential loss of eukaryotic genes. *Nature* **524**, 427–432 (2015).
- Lynch, M. & Conery, J. S. The origins of genome complexity. *Science* **302**, 1401–1404 (2003).
- Shapiro, B. J. How clonal are bacteria over time? *Curr. Opin. Microbiol.* **31**, 116–123 (2016).
- Vos, M., Hesselman, M. C., te Beek, T. A., van Passel, M. W. & Eyre-Walker, A. Rates of lateral gene transfer in prokaryotes: high but why? *Trends Microbiol.* **23**, 598–605 (2015).
- Kimura, M. *The Neutral Theory of Molecular Evolution* (Cambridge Univ. Press, 1984).
- Lane, N. & Martin, W. The energetics of genome complexity. *Nature* **467**, 929–934 (2010).
- Ohta, T. Slightly deleterious mutant substitutions in evolution. *Nature* **246**, 96–98 (1973).
- Konstantinidis, K. T. & Tiedje, J. M. Trends between gene content and genome size in prokaryotic species with larger genomes. *Proc. Natl Acad. Sci. USA* **101**, 3160–3165 (2004).
- Kuo, C. H. & Ochman, H. Deletional bias across the three domains of life. *Genome Biol. Evol.* **1**, 145–152 (2009).
- Sela, I., Wolf, Y. I. & Koonin, E. V. Theory of prokaryotic genome evolution. *Proc. Natl Acad. Sci. USA* **113**, 11399–11407 (2016).
- Nakamura, Y., Itoh, T., Matsuda, H. & Gojobori, T. Biased biological functions of horizontally transferred genes in prokaryotic genomes. *Nat. Genet.* **36**, 760–766 (2004).
- Pandey, D. P. & Gerdes, K. Toxin-antitoxin loci are highly abundant in free-living but lost from host-associated prokaryotes. *Nucleic Acids Res.* **33**, 966–976 (2005).
- McNally, A. *et al.* Combined analysis of variation in core, accessory and regulatory genome regions provides a super-resolution view into the evolution of bacterial populations. *PLoS Genet.* **12**, e1006280 (2016).
- Baltrus, D. A. Exploring the costs of horizontal gene transfer. *Trends Ecol. Evol.* **28**, 489–495 (2013).
- Charlesworth, B. Fundamental concepts in genetics: effective population size and patterns of molecular evolution and variation. *Nat. Rev. Genet.* **10**, 195–205 (2009).
- Niehus, R., Mitri, S., Fletcher, A. G. & Foster, K. R. Migration and horizontal gene transfer divide microbial genomes into multiple niches. *Nat. Commun.* **6**, 8924 (2015).
- Karcagi, I. *et al.* Indispensability of horizontally transferred genes and its impact on bacterial genome streamlining. *Mol. Biol. Evol.* **33**, 1257–1269 (2016).
- Hutchison, C. A. 3rd *et al.* Design and synthesis of a minimal bacterial genome. *Science* **351**, aad6253 (2016).
- Chang, Y. J. *et al.* Non-contiguous finished genome sequence and contextual data of the filamentous soil bacterium *Ktedonobacter racemifer* type strain (SOSP1–21). *Stand. Genomic Sci.* **5**, 97–111 (2011).
- Lee, M. C. & Marx, C. J. Repeated, selection-driven genome reduction of accessory genes in experimental populations. *PLoS Genet.* **8**, e1002651 (2012).
- Locey, K. J. & Lennon, J. T. Scaling laws predict global microbial diversity. *Proc. Natl Acad. Sci. USA* **113**, 5970–5975 (2016).
- Erwin, D. H. A public goods approach to major evolutionary innovations. *Geobiology* **13**, 308–315 (2015).
- McInerney, J. O., Pisani, D., Bapteste, E. & O’Connell, M. J. The public goods hypothesis for the evolution of life on Earth. *Biol. Direct* **6**, 41 (2011).
- Schatz, M. C. *et al.* Whole genome *de novo* assemblies of three divergent strains of rice, *Oryza sativa*, document novel gene space of *aus* and *indica*. *Genome Biol.* **15**, 506 (2014).
- Li, Y. H. *et al.* *De novo* assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits. *Nat. Biotechnol.* **32**, 1045–1052 (2014).
- Read, B. A. *et al.* Pan genome of the phytoplankton *Emiliania underpins* its global distribution. *Nature* **499**, 209–213 (2013).
- Ding, W., Baumdicker, F. & Neher, R. A. panX: pan-genome analysis and exploration. Preprint at *bioRxiv* <https://doi.org/10.1101/072082> (2016).
- Sharp, P. M., Stenico, M., Peden, J. F. & Lloyd, A. T. Codon usage: mutational bias, translational selection, or both? *Biochem. Soc. Trans.* **21**, 835–841 (1993).
- McInerney, J. O. Replicational and transcriptional selection on codon usage in *Borrelia burgdorferi*. *Proc. Natl Acad. Sci. USA* **95**, 10698–10703 (1998).
- McInerney, J. O. Prokaryotic genome evolution as assessed by multivariate analysis of codon usage patterns. *Microb. Comp. Genomics* **2**, 89–97 (1997).
- Doherty, A. & McInerney, J. O. Translational selection frequently overcomes genetic drift in shaping synonymous codon usage patterns in vertebrates. *Mol. Biol. Evol.* **30**, 2263–2267 (2013).

Acknowledgements

We wish to thank J. Mallet for commenting on a draft of this manuscript. We would also like to thank the anonymous reviewers. J.O.M. is funded by BBSRC grant no. BB/N018044/1 and the John Templeton Foundation.

Author contributions

J.O.M., A.M. and M.J.O. collectively conceived and wrote this manuscript.

Additional information

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to J.O.M.

How to cite this article: McInerney, J. O., McNally, A. & O’Connell, M. J. Why prokaryotes have pangenomes. *Nat. Microbiol.* **2**, 17040 (2017).

Publisher’s note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Competing interests

The authors declare no competing financial interests.