UNIVERSITY OF LEEDS

This is a repository copy of *Why Prokaryotes Have Pangenomes*.

White Rose Research Online URL for this paper:
https://eprints.whiterose.ac.uk/id/eprint/113972/

Version: Accepted Version

**Article:**
McInerney, JO, McNally, A and O'Connell, MJ orcid.org/0000-0002-1877-1001 (2017) Why Prokaryotes Have Pangenomes. Nature Microbiology, 2 (4). 17040. ISSN: 2058-5276

https://doi.org/10.1038/nmicrobiol.2017.40

© 2017 Macmillan Publishers Limited, part of Springer Nature. This is an author produced version of a paper published in Nature Microbiology. Uploaded in accordance with the publisher's self-archiving policy. The final publication is available at https://doi.org/10.1038/nmicrobiol.2017.40

**Reuse**
Items deposited in White Rose Research Online are protected by copyright, with all rights reserved unless indicated otherwise. They may be downloaded and/or printed for private study, or other acts as permitted by national copyright laws. The publisher or other rights holders may allow further reproduction and re-use of the full text version. This is indicated by the licence information on the White Rose Research Online record for the item.

**Takedown**
If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

White Rose university consortium
Universities of Leeds, Sheffield & York

eprints@whiterose.ac.uk
https://eprints.whiterose.ac.uk/

# Why Prokaryotes Have Pangenomes.

James O. McInerney[1]*, Alan McNally[2] and Mary J. O'Connell[3]

[1]Division of Evolution and Genomic Sciences, School of Biological Sciences, Faculty of Biology, Medicine and Health, The University of Manchester, Manchester Academic Health Science Centre, Manchester M13 9PL, United Kingdom.

[2]Institute of Microbiology and Infection, College of Medical and Dental Sciences, University of Birmingham, Edgbaston, Birmingham B15 2TT, United Kingdom.

[3]Computational and Molecular Evolutionary Biology Group, School of Biology, Faculty of Biological Sciences, The University of Leeds, LS2 9JT, United Kingdom

| | |
|---|---|
| Email: | james.mcinerney@manchester.ac.uk |
| Twitter: | @McInerneyLab |
| Facebook: | /McInerneyLab |
| Web: | http://McInerneyLab.com/ |

**Summary**

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 the focal species has the ability to migrate to new niches. Therefore, we suggest that pangenomes are the result of adaptive, not neutral evolution.

**Introduction**

It became apparent as soon as different strains of the same species had their genomes sequenced that there was enormous intraspecific variability in prokaryotic genome content[1]. Indeed, terms such as "core" and "accessory" genome, have been coined in order to describe this variation[2]. 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 genome[3]. The "pangenome" consists of all the gene families that have been found in the species as a whole[4] (see figure 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 one [5]. However, the absence of theory to explain pangenomes is a gap in the New Synthesis. In this paper 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 HGT did not affect phylogenies [9].

Today the thousands of prokaryotic genome sequences available reveal the pervasive influence of introgressions of many kinds[10]. The largest pangenome analysis for a single species to date included 2,085 *E. coli* genomes[11] which estimated 3,188 core gene families (which they defined as being present in 95% of genomes) 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 bigger[12], 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 of genomes, the entire bacterial pangenome has been estimated to be infinite in size[13]. 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 prokaryotes[15] and higher levels of gene duplication [16]. In this paper 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 plenitude of mobile genetic elements and exogenous DNA that gene delivery is quite frequent; the question is what promotes retention and why there isn't 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 text box 1 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.

**Text Box 1 - Small Selective Pressures**

Selection for mutations that confer even very small fitness effects can be seen in organisms with large $N_e$[18]. *Escherichia coli* has a large pangenome and $N_e$ in this species is estimated to be 25,000,000[18]. 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 tRNAs[19]. 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 genes[19]. *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 selection[20], as well as selection for function[21] and the cost of maintenance[22]. Not all organisms have very large $N_e$, however. In particular pathogens or symbionts that frequently encounter bottlenecks during transmission have small $N_e$[18]. The obligately intracellular pathogen, *Mycoplasma genitalium*, which likely has a small $N_e$, does not show evidence of translational selection in any genes, no matter whether expressed at high or low levels[23]. In the human genome, drift is not overcome by selection for codon usage optimisation[24]. 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

**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 allele[25]. 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 $\mu$ 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 is equal to the rate of introduction of the alleles[26]. 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 translated[22]. 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 $|s| << 1/N$, *i.e.* a nearly neutral allele, *sensu* Ohta[27]. 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[28]. Indeed empirical genome analyses have demonstrated that prokaryotic genomes are biased towards deletion of DNA [29], 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 adaptive[30].

**Models with associated fitness costs.**

Another potential explanation for pangenomes is that accessory genes are composed largely of selfish or addictive genetic elements and the existence of extreme genome variability is 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 scenario[31]. Of course, some accessory genes are selfish elements such as phage or toxin-antitoxin genes [32], but thousands of known accessory genes have other known functions[28] and do not appear to have "addictive" traits, so a theory based on selfish genes is insufficient here (see Figure 2 for accessory gene analysis of 228 *E. coli* ST131 genomes) [33].

Comparison of closely-related genomes indicates that many HGTs are relatively transient, being frequently supplanted by other newcomers[12]. 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 as well as the system-level effects of having additional protein products in a cell[34]. However, while HGT can have these costs, if HGT were 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 rates[25]. It is clear that HGT rates, at least in some organisms, are quite high[30], 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 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 recombination[18]. 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[35] 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 from 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 species[35]. 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 like *Chlamyda 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 text box 1), selection overcoming drift in prokaryotes is crucially dependent on the $N_e$ 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 [21,36]. 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 HGT[21]. 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. [36] 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 don't 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 sizes[37], 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 Marx[38] have shown selection-driven genome reduction in *Methylobacterium extorquens* AM1 experimental populations. Further investigation revealed a "decreased performance" of reduced-genome *Mb. 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, 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 1 Trillion ($10^{12}$) microbial species on Earth[39], 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[40,41]. Future empirical work will involve understanding the precise interplay between HGT, selection, drift, migration, population size, and pangenomes.

Figure Legends:

**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.

**Figure 2**: Analysis of accessory gene functions in 228 *Escherichia coli* ST131 genomes. Though selfish elements constitute a large portion of the known functions, they are not the majority.

**Acknowledgements**: We wish to thank James Mallet for commenting on a draft of this manuscript. We would also like to thanks the anonymous reviewers. JOM is funded by BBSRC grant No. BB/N018044/1 and the John Templeton Foundation.

**References:**

1 Perna, N. T. *et al.* Genome sequence of enterohaemorrhagic Escherichia coli O157:H7. *Nature* **409**, 529-533 (2001).
2 Young, J. P. *et al.* The genome of Rhizobium leguminosarum has recognizable core and accessory components. *Genome Biol* **7**, R34 (2006).
3 Tettelin, H. *et al.* Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome". *Proc Natl Acad Sci U S A* **102**, 13950-13955 (2005).
4 Ku, C. *et al.* Endosymbiotic gene transfer from prokaryotic pangenomes: Inherited chimerism in eukaryotes. *Proc Natl Acad Sci U S A* (2015).
5 Treangen, T. J. & Rocha, E. P. Horizontal transfer, not duplication, drives the expansion of protein families in prokaryotes. *PLoS Genet* **7**, e1001284 (2011).
6 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).
7 Creevey, C. J. *et al.* Does a tree-like phylogeny only exist at the tips in the prokaryotes? *Proceedings of the Royal Society of London. Series B: Biological Sciences* **271**, 2551-2558 (2004).
8 Doolittle, W. F. Phylogenetic classification and the universal tree. *Science* **284**, 2124-2129 (1999).
9 Daubin, V., Moran, N. A. & Ochman, H. Phylogenetics and the cohesion of bacterial genomes. *Science* **301**, 829-832 (2003).
10 Bapteste, E. *et al.* Evolutionary analyses of non-genealogical bonds produced by introgressive descent. *Proc Natl Acad Sci U S A* **109**, 18266-18272 (2012).
11 Land, M. *et al.* Insights from 20 years of bacterial genome sequencing. *Funct Integr Genomics* **15**, 141-161 (2015).
12 Lukjancenko, O., Wassenaar, T. M. & Ussery, D. W. Comparison of 61 sequenced Escherichia coli genomes. *Microbial ecology* **60**, 708-720 (2010).
13 Lapierre, P. & Gogarten, J. P. Estimating the size of the bacterial pan-genome. *Trends Genet* **25**, 107-110 (2009).
14 Li, R. *et al.* Building the sequence map of the human pan-genome. *Nat Biotechnol* **28**, 57-63 (2010).
15 Ku, C. *et al.* Endosymbiotic origin and differential loss of eukaryotic genes. *Nature* **524**, 427-432 (2015).
16 Lynch, M. & Conery, J. S. The origins of genome complexity. *Science* **302**, 1401-1404 (2003).
17 Shapiro, B. J. How clonal are bacteria over time? *Curr Opin Microbiol* **31**, 116-123 (2016).
18 Charlesworth, B. Fundamental concepts in genetics: effective population size and patterns of molecular evolution and variation. *Nat Rev Genet* **10**, 195-205 (2009).
19 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).
20 McInerney, J. O. Replicational and transcriptional selection on codon usage in Borrelia burgdorferi. *Proc Natl Acad Sci U S A* **95**, 10698-10703 (1998).
21 Karcagi, I. *et al.* Indispensability of Horizontally Transferred Genes and Its Impact on Bacterial Genome Streamlining. *Mol Biol Evol*, doi:10.1093/molbev/msw009 (2016).
22 Lane, N. & Martin, W. The energetics of genome complexity. *Nature* **467**, 929-934 (2010).

23 McInerney, J. O. Prokaryotic genome evolution as assessed by multivariate analysis of codon usage patterns. *Microbial & Comparative Genomics* **2**, 89-97 (1997).
24 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).
25 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).
26 Kimura, M. *The neutral theory of molecular evolution*. (Cambridge University Press, 1984).
27 Ohta, T. Slightly deleterious mutant substitutions in evolution. *Nature* **246**, 96-98 (1973).
28 Konstantinidis, K. T. & Tiedje, J. M. Trends between gene content and genome size in prokaryotic species with larger genomes. *Proc Natl Acad Sci U S A* **101**, 3160-3165 (2004).
29 Kuo, C. H. & Ochman, H. Deletional bias across the three domains of life. *Genome Biol Evol* **1**, 145-152 (2009).
30 Sela, I., Wolf, Y. I. & Koonin, E. V. Theory of prokaryotic genome evolution. *Proc Natl Acad Sci U S A* **113**, 11399-11407 (2016).
31 Nakamura, Y., Itoh, T., Matsuda, H. & Gojobori, T. Biased biological functions of horizontally transferred genes in prokaryotic genomes. *Nat Genet* **36**, 760-766 (2004).
32 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).
33 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).
34 Baltrus, D. A. Exploring the costs of horizontal gene transfer. *Trends Ecol Evol* **28**, 489-495 (2013).
35 Niehus, R., Mitri, S., Fletcher, A. G. & Foster, K. R. Migration and horizontal gene transfer divide microbial genomes into multiple niches. *Nature communications* **6**, 8924 (2015).
36 Hutchison, C. A., 3rd *et al.* Design and synthesis of a minimal bacterial genome. *Science* **351**, aad6253 (2016).
37 Chang, Y. J. *et al.* Non-contiguous finished genome sequence and contextual data of the filamentous soil bacterium Ktedonobacter racemifer type strain (SOSP1-21). *Standards in genomic sciences* **5**, 97-111 (2011).
38 Lee, M. C. & Marx, C. J. Repeated, selection-driven genome reduction of accessory genes in experimental populations. *PLoS Genet* **8**, e1002651 (2012).
39 Locey, K. J. & Lennon, J. T. Scaling laws predict global microbial diversity. *Proc Natl Acad Sci U S A* **113**, 5970-5975 (2016).
40 Erwin, D. H. A public goods approach to major evolutionary innovations. *Geobiology*, doi:10.1111/gbi.12137 (2015).
41 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).
42 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).
43 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).
44 Read, B. A. *et al.* Pan genome of the phytoplankton Emiliania underpins its global distribution. *Nature* **499**, 209-213 (2013).
45 Ding, W., Baumdicker, F. & Neher, R. A. *panX: pan-genome analysis and exploration* (Biorxiv, 2016).

Figure 1:

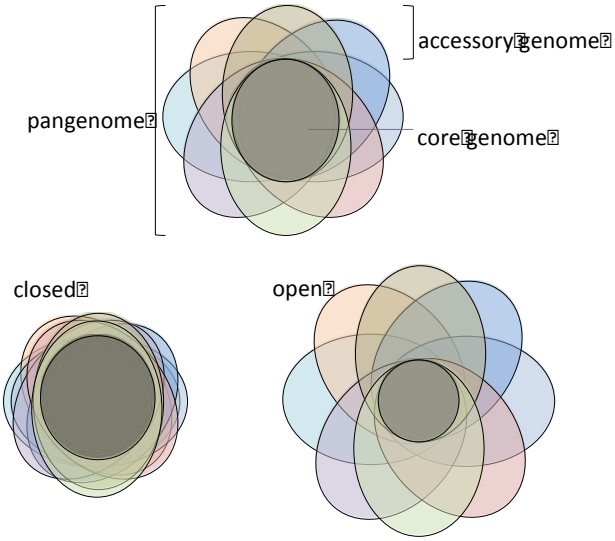
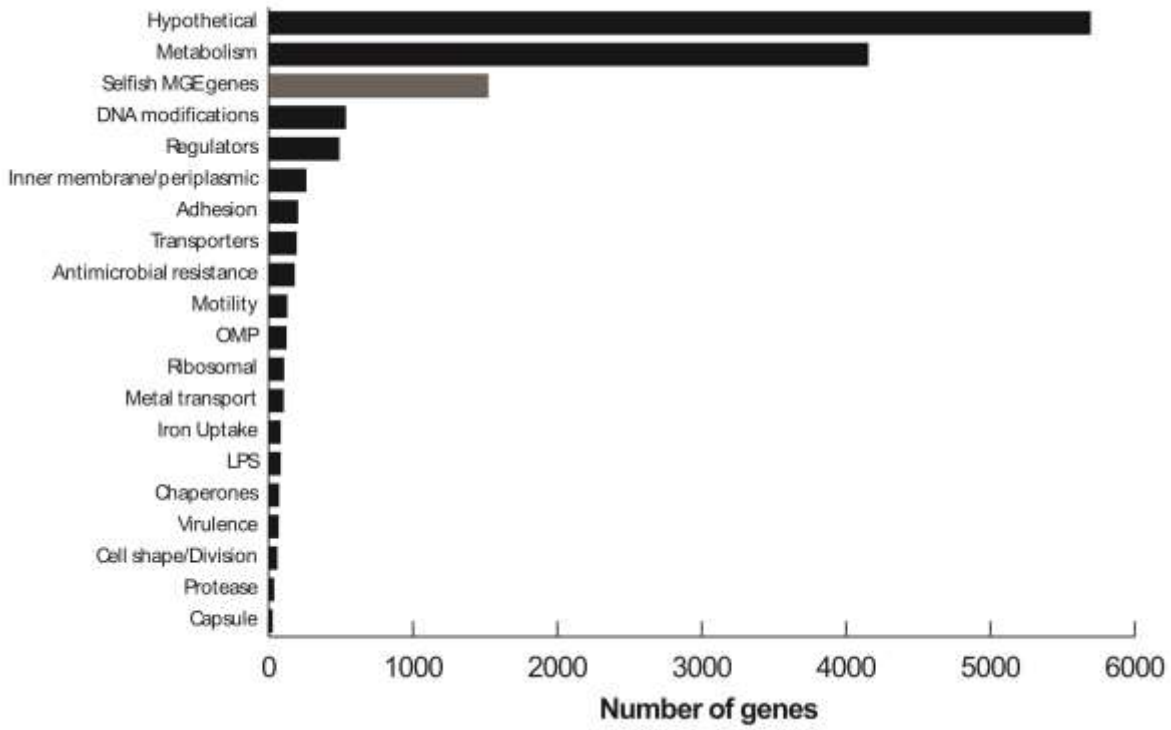



Figure 2:




**Glossary box**

**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 (HGT):** 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 observation that explains 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**: This is 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**: These are 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 the antitoxin being depleted 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**: This refers to 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.