Structure and Dynamics of Bacterial Populations: Pangenome Ecology

Taj Azarian, I-Ting Huang, and William P. Hanage

Abstract Prokaryotes demonstrate tremendous variation in gene content, even within individual bacterial clones or lineages. This diversity is made possible by the ability of bacteria to horizontally transfer DNA through a variety of mechanisms, and the extent of such transfer sets them apart from eukaryotes. What has become evident through interrogation of thousands of bacterial genomes is that gene variation is directly related to the ecology of the organism and is driven by continual processes of niche exploration, diversification, and adaptation. Of course, the acquisition of new genes is not necessarily beneficial, resulting in either the removal of that individual through purifying selection or the occurrence of compensatory mutations in the genomic “backbone” (i.e., core genes) that become epistatically linked to the presence accessory genes. There are now numerous examples of relationship between gene variation and niche adaptation. We explore some of those examples here as well as the population genomic footprint left by the dynamics of gene flow, diversification, and adaptation.

Keywords Bacterial population genomics · Pangenome · Accessory genome · Ecology · Adaptation · Bacterial evolution · Recombination · Horizontal gene transfer

T. Azarian

College of Medicine, University of Central Florida, Orlando, FL, USA

I.-T. Huang · W. P. Hanage (✉)

Center for Communicable Disease Dynamics, Department of Epidemiology, T.H. Chan School of Public Health, Harvard University, Boston, MA, USA

e-mail: whanage@hsph.harvard.edu

1 Introduction

Pangenomes of bacterial species show a tremendous range of diversity in size, content, and fluidity. In comparing the core genome size in relation to the accessory genome, some species possess relatively limited pangenomes while others are expansive. Accessory genomes may be composed of genes belonging to phages, transposons, insertion sequences, and plasmids, as well as genes that have diverged through mutation and recombination to the point where they are considered as a separate homolog. Some of these genomic elements may be relatively stable (e.g., an integrated prophage), while others may be gained and lost within a single bacterial culture (e.g., plasmids). In this chapter, we will discuss the population genomics of pan-, and more specifically, accessory genomes, specifically detailing how accessory genomes vary among and within bacterial species and the implications this variation has for microbial ecology. Throughout this discussion, it is important to not lose sight of what we are referring to with the catch-all phrase “accessory.” These are the dynamic elements of the genome, often containing large genomic islands that augment the bacterium’s phenotype, which may, as we will outline, be used to glean knowledge of ecology and evolutionary history of a genus, species, or set of lineages. Further, in no way does the term accessory or the misleading synonym “dispensable” suggest non-essential, as some “accessory” genes actually represent divergent variants of an essential gene.

2 Mechanisms of Pangenome Variation

The content and diversity of a bacteria’s accessory genome are directly associated with the mode and frequency of horizontal gene transfer (HGT), which in turn is tightly linked to ecology. Modes of HGT include transformation: the uptake and integration of exogenous DNA from the environment, transduction: the introduction of exogenous DNA into the bacterial cell through a viral vector (e.g., bacteriophage), and conjugation: the direct transfer of DNA between two bacterial cells through a pilus, which usually involves plasmids and transposons. Bacteria vary in the degree to which each of these mechanisms occurs within their populations and in their DNA uptake mechanisms. It is also almost certain that other variants of these mechanisms remain to be discovered, as illustrated by recent work describing “lateral transduction” capable of transferring genomic regions of remarkable size (Chen et al. 2018).

Integrative and conjugative elements (ICE) include integrative plasmids and conjugative transposons, which are circularized mobile elements transferred through conjugation. ICE may harbor a number of genes important to virulence, specialized metabolism, and survival, and are the primary means by which antibiotic-resistant genes are transmitted among bacteria. Plasmids may contain anywhere from 5 to 100 or so genes, allowing for a lineage to gain or lose many loci in a single step, especially for those species with high plasmid diversity. Phylum *Proteobacteria*,

which includes several pathogenic species from genera *Escherichia*, *Salmonella*, *Vibrio*, *Helicobacter*, *Yersinia*, and *Legionellales* possess some of the most prevalent and diverse plasmids with a wide host range (Shintani et al. 2015). Therefore, unsurprisingly, species among these genera have moderately large pangenome sizes (McInerney et al. 2017).

Naturally, competent (transformable) species are able to uptake DNA directly from the environment resulting in homologous or nonhomologous recombination, the latter frequently associated with gene gain (Croucher et al. 2012). Arguably, the most famous of these species, *Streptococcus pneumoniae*, was made so by its role in the Griffith experiments in 1928, which led to the identification of DNA as the conveyor of genetic information. Through those experiments, Griffith observed that “smooth” (i.e., unencapsulated) avirulent *S. pneumoniae* could become virulent through exposure to heat-killed virulent “rough” (i.e., encapsulated) pneumococci (Griffith 1928). We now know that what he observed was transformation resulting in the acquisition of the capsular polysaccharide (CPS) loci that code genes responsible for the synthesis and polymerization the antigenic serotype capsule. There are over 90 serotypes identified and the CPS loci span 10,337–30,298 bp with at least 26 coding sequences depending on the particular serotype (Bentley et al. 2006). Therefore, this single recombination event resulted in the acquisition of 26 accessory genes. Since then, other species including *Neisseria gonorrhoeae*, *Campylobacter jejuni*, *Vibrio cholerae*, and *Haemophilus influenzae* have been found to be naturally competent.

Another method by which transformation may result in differences in gene content is through events that lead to gene diversification, which are frequently observed among several species as recombination “hotspots.” The primary effect of these events is antigenic variation in genes linked to host–pathogen interactions. For example, among pneumococci, two virulence factors, pneumococcal surface proteins A and C (pspA and pspC), are known to have 3 and 11 variants, respectively (Hollingshead et al. 2000; Iannelli et al. 2002). These variants are diverse in length, structural organization, and nucleotide variation, the results of frequent recombination events. Most important, they are different in serology, which has significant implications for host immunity (Azarian et al. 2016; Georgieva et al. 2018). Similarly, among gonococci, the *opa* and neighboring *pil* loci are highly mosaic due to recombination of existing alleles (Bilek et al. 2009). The gene product *Opa* is an outer membrane adhesion protein that is important for colonization and invasion of the genital and nasopharyngeal mucosal epithelium. As a note, antigenic variation through recombination leads to an interesting contradiction in terminology. In both of these examples, pspA, pspC, *opa*, and *pil* are considered “core” genes in the sense that each member of their respective species possesses a variant. They are by all definitions “essential” to core cell function; yet, through current methods of pangenome analysis that are commonly based on a nucleotide homology level of at least 80%, they are identified operationally as accessory genes. Finally, transduction through temperate bacteriophages may introduce considerable gene variation in both Gram-negative and Gram-positive bacteria (Feng et al. 2008; Waldor and Friedman 2005). While their precise evolutionary impact in most cases remains unclear, it is

certain that their pathogenesis plays a significant role in the biology of their host. For example, many phages harbor genes coding for virulence factors including toxins or secreted enzymes (Romero et al. 2009); therefore, prophages (bacteriophages integrated into host bacterial genomes) represent a significant mechanism for variation of virulence among closely related bacteria (Fortier and Sekulovic 2013). In relation to pangenome dynamics, the transmission of bacteriophages can result in significant variation among bacterial populations on short timescales by two mechanisms: through (1) the direct integration of the prophage and (2) the acquisition or evolution of antiphage mechanisms. The later may involve phage-inducible chromosomal islands and CRISPR-Cas systems (Reyes-Robles et al. 2018), which independently represent instances of gene acquisition and a source of pangenome variation. Predator–prey dynamics of bacteriophages and their host has been widely observed with *Siphoviridae* phages and *S. pneumoniae* (Romero et al. 2009), lambda STX-coding phage in Shiga toxin-producing *E. coli*, and ICP (*Myoviridae*) and CTX phages in *Vibrio cholerae* (Seed et al. 2011; Waldor and Friedman 2005), among myriad others. The result is highly variable prophage content even within closely related members of bacterial lineages (Croucher et al. 2014).

3 Population Genomics of Pangenomes

Today, the identification of a bacterial sample's core genome is a common intermediate step among bioinformatics pipelines for preparing whole-genome sequencing data for phylogenetic analysis. Historically, the accessory genome was largely ignored with the exception of the identification of important genes such as those conferring antibiotic resistance or increased virulence. Methodologically, it was difficult to scale accessory genome analysis to large population samples of a species and especially across several species. Then, the discovery that in three diverse *E. coli* isolates, less than 40% of the genes was found in the genomes of all three demonstrated that extensive variation was possible (Welch et al. 2002). A subsequent study of just eight genomes of *Streptococcus agalactiae* (Group B *Streptococcus*) published in 2005 identified 1806 core genes and 439 “dispensable” genes, highlighting that tremendous variation could be observed with even a small sample (Tettelin et al. 2005). This chapter introduced the concept of the pangenome. Now, large-scale analyses of pangenomes continue to reveal significant diversity even over short timescales, providing information about the demographic history and adaptive evolution of bacteria. These studies have shown that pangenome size and diversity vary among species and depend on lifestyle (McInerney et al. 2017; Ochman and Davalos 2006).

McInerney and colleagues recently summarized the range of diversity observed among bacterial species (McInerney et al. 2017). Pangenome sizes ranged from 974 for the obligate intracellular bacteria *Chlamydia trachomatis* to 40,362 for the semiaquatic agricultural *Oryza sativa*. Comparing sizes of accessory genomes in relation to the total number of genes in the pangenome, *O. sativa* had the smallest,

just 8% of genes were accessory, while in *Salmonella enterica*, a staggering 83% of its 10,267 genes are found in the accessory genome. Assessing “genomic fluidity” is another method for quantifying pangenome diversity (Kislyuk et al. 2011). Instead of assessing the relationship between core and accessory genome size, genomic fluidity measures the dissimilarity of genomes evaluated at the gene level calculated as the “ratio of unique gene families to the sum of gene families in pairs of genomes averaged over randomly chosen genome pairs from within a group of N genomes.” In a comparison of genomic fluidity among seven species known to undergo HGT, *Neisseria meningitidis*, *Escherichia coli*, and *Streptococcus* spp. ranked highest in genomic fluidity (Kislyuk et al. 2011) (although it should be noted that this metric is expected to be affected by the sample chosen for study).

Within a species, accessory genome diversity increases with core genome divergence and models of homologous recombination and HGT have shown how these processes lead to the formation of population structure (Croucher et al. 2014; Marttinen et al. 2015). Boundaries for HGT across species roughly follow the same trajectories. Species in genera *Streptococcus*, *Neisseria*, and *Campylobacter*, for example, have been shown to engage in HGT more frequently with closely related members (e.g., between *S. pneumoniae* and *S. mitis* and *S. oralis*, and *N. gonorrhoeae* and *N. meningitidis*). Therefore, the size and distribution of accessory genes in a population provide insights into the demographic history of bacterial species as well as delineations of species boundaries.

As we have described, many methods can generate accessory genome diversity. While not wholly analogous to the way nucleotide mutations arise and propagate in a population, the gain and loss of genes nonetheless inform the shared evolutionary history of a population in the same manner. Genomic islands acquired through HGT often become relatively fixed in bacterial lineages (Croucher et al. 2014) with the number of acquired genes increasing with lineage age (Donati et al. 2010). This is especially true for Staphylococcal Cassette Chromosome *mec* (*SCCmec*) elements in clones of *S. aureus* (International Working Group on the Classification of Staphylococcal Cassette Chromosome Elements (IWG-SCC) 2009), pathogenicity islands among toxigenic and non-toxigenic lineages of *V. cholerae* (Wozniak et al. 2009), and CPS loci in pneumococci (Bentley et al. 2006). These mobile elements, therefore, inform long-scale evolutionary history, while in the short term, prophage variation and the scars of transformation events reflect more recent events. As such, it is possible to recapitulate the core genome phylogeny of a population through phylogenetic reconstruction using a presence–absence alignment of accessory genes, represented by 1’s and 0’s, respectively (Azarian et al. 2018). In essence, this represents a tight linkage between core genome single nucleotide polymorphisms and the history of gene gain and loss. This may, of course, oversimplify the complex interconnected processes that led to accessory gene variation, but it does provide an easy data structure that may be investigated to understand how bacterial populations change over time.

An interesting approach to assessing temporal changes in bacterial population genomics is to consider the dynamics of the accessory genome. The clearest examples of this are observations of rapid changes in virulence or antibiotic

resistance among bacterial lineages, often leading to short-term success of a clone (Croucher et al. 2014). The impact of human interventions, namely vaccines, affects not only the distribution of lineages in a population but also the available pool of accessory genes. For example, if an ICE is strongly associated with a lineage, and that lineage is targeted by vaccine, then the removal of the lineage from the population may ultimately remove the reservoir for that ICE. The impact of vaccine on the pathogen population of *S. pneumoniae* has been extensively studied (Azarian et al. 2018; Croucher et al. 2013). After the introduction of the seven-valent pneumococcal conjugate vaccine (PCV7) in the USA, an analysis of a sample of 616 genomes of pneumococci carried in children in Massachusetts showed the removal of accessory genes associated with the CPS loci of vaccine serotypes (Croucher et al. 2013). In addition, the prevalence of antibiotic-resistant genes associated with two transposons was shifted due to the removal of two vaccine lineages they were associated with and the subsequent emergence of a non-vaccine lineage harboring one of the transposons. A study of pneumococcal population dynamics over 13 years and spanning the introduction of the PCV7 showed that the introduction of vaccines greatly shifted the frequencies of accessory genes in the population (Azarian et al. 2018). Surprisingly, the frequencies of accessory genes then shifted back to pre-vaccine values as the pneumococcal population recovered from the removal of nearly 30% prevalent genotypes targeted by vaccine. This observation was elucidated by recent work by Corander and colleagues who investigated accessory gene frequencies across of 4127 pneumococcal isolates from four distinct geographic areas (Corander et al. 2017). They found that accessory genes had similar frequencies in the four populations despite significant differences in lineage composition and the timing of vaccine use. Through functional analysis of the accessory genes and population dynamic modeling, they proposed that the frequencies of accessory genes are shaped by negative frequency-dependent selection (NFDS) through pathogen–pathogen, host–pathogen, and pathogen–environment interactions. Classically defined, in an NFDS model the fitness of a phenotype depends on its frequency relative to other phenotypes in a population. The same NFDS model has been used to explain the diversity of protein antigens among pneumococci, which we briefly touched upon early in the chapter. In the case of protein antigens, increasing host immunity toward an antigen drives diversification of the gene coding for the protein either through mutation, or most often, recombination. The same dynamic can be observed with prophages and restriction modification systems that defend against infection. Ultimately, these observations point to a central hypothesis for accessory genome variation, that difference in gene content are linked to adaptation and niche specialization, but that in the case of NFDS the niche may be dynamically generated by fluctuating frequencies of loci in the pangenome.

4 The Ecological Significance of Pangenomes

The observation of pangenomes as a common feature of many bacteria begs the question of what has selected them? What are the ecological features that lead to the pervasive association of a core, with a disseminated complement of many additional genes, some shared with other species? While some have clear selective consequences, most are obscure. The extent to which bacteria vary in gene content sets them apart from eukaryotes, and is just one of the reasons we cannot easily transfer population genetic concepts between the superkingdoms of life. One metaphor for bacteria and their varying genome content compares them to modern smartphones (Young 2016) in which the core genome is the operating system, the accessory genome is the apps downloaded to the phone, and the pangenome would be everything in the app store. In the following, we divide up the accessory genes that combine to make up a pangenome into various categories, not by function but by how they are distributed among lineages in the population.

The perspective we take is of the bacterial genome as a transient construct. Loci can be added to it, and selected to become more common or indeed lost from the population, should they no longer be necessary. The pangenome for any sample is the totality of genes currently associated with its contents. This need not be a permanent or even especially long relationship. Consider a locally prominent prophage, which might not be present in the same population if you returned at a later date. Indeed we can imagine that given the many ways bacteria engage in HGT, a sample of sufficient size will contain many loci in a new genetic background that are yet to be lost (analogous to incomplete purifying selection (Rocha et al. 2006). A subset of the pangenome, expected to be rare in any reasonably large sample, is genes that are either infrequently obtained or actively selected against. In general, the extent of gene flow will be regulated by the genetic and ecological similarity of the bacteria and the compatibility of the genetic background to adapt to the acquisition of novel genes (Wiedenbeck and Cohan 2011).

Moving to loci that are present at intermediate frequencies, say between 5% and 95% of isolates, we can distinguish between loci that are restricted to a few lineages, or are widely disseminated but not fixed in any lineage. These suggest different evolutionary scenarios. Dealing with the latter first; a locus that is easy to obtain but hard to hold onto suggests fluctuating selection. We see it more often than the genes in the previous category, because it provides selective benefits. However, these are not consistent benefits or we would expect the gene to rapidly become more common and indeed part of the core. Examples of these include drug-resistant genes in lineages that lack compensatory mutations, and as such only experience a selective benefit in the presence of the drug (Blanquart et al. 2018; Cobey et al. 2017; Lehtinen et al. 2017).

In contrast, loci fixed in a lineage might represent the ecological “address” of those bacteria, a dimension of their niche. However, this need not be the case. Studies of populations of *S. pneumoniae* have shown that the accessory loci in this species are not widely disseminated, but are also rarely restricted to a single lineage

and are instead shared among several, in different combinations (Croucher et al. 2014). It has been suggested that different combinations of accessory loci might be selected in different populations, depending on the overall frequencies of the individual genes, as a result of negative frequency-dependent selection (Corander et al. 2017). At present, this remains a hypothesis without definitive proof.

We must also recognize that a locus might have no wider ecological significance whatsoever. Toxin–antitoxin genes can drive their own acquisition and maintenance, to say nothing of the multitudes of transposable elements, prophage and the like (Wozniak and Waldor 2009). Bacterial genomes are characterized not only by their variable gene content, and the transience of the associations between loci (long for core genes, short for others) but by the divergent selective processes affecting them. In some cases (the core) these are aligned, while in others they are not. Population geneticists who study sexually reproducing eukaryotes are familiar with the notion that the selective interests of different loci in the same genome may differ. The shuffling of genetic information in each generation effectively uncouples the association between all but the closest loci, but even the most frequently recombining bacteria (Arnold et al. 2018) do not approach the state of sexually reproducing eukaryotes. As a result, the overall fitness of a bacterial genome is the product of all the loci making it up. To preserve this overall fitness, it has been proposed that homologous recombination in bacteria is an adaptation to prevent the colonization of the genome with selfish genetic elements, by rapidly replacing them with the homologous region in the ancestral strain, which lacks the additional gene (although this does not explain the notable variation among bacteria in their recombination rates) (Croucher et al. 2016). One of the greatest challenges in providing a satisfying account of bacterial population genetics has been separating the patterns that are the result of selection, from those of linkage.

The question of how the individual loci that make up the accessory or dispensable part of the pangenome, associate themselves with the lineages that are defined by the core component, has come under increasing scrutiny as the numbers of population genome samples have increased. Population genetic models for the core genome specifically developed with bacteria in mind, and capable of handling the various amounts of homologous recombination, are not common. Rarer still are models that explicitly consider the gain and loss of genes from the accessory genome. Although gain and loss of loci is not unknown in eukaryotes, and has been implicated in some major adaptive events (McInerney 2017; Schönknecht et al. 2014) it is nowhere near as extensive and does not have anything like the impact it does in bacteria. Accurate models for such processes are crucial to detect departures from neutrality, and several studies have actually found apparently neutral associations between elements of the pangenome and the core. However there are reasons to think that the sequence variation associated with the accessory genome may produce fundamentally different results from those in population genetics textbooks. For example, if the site frequency spectrum expected under neutral assumptions is extended to allow mutations in loci that can be gained or lost, systematic bias results (Baumdicker 2015; Baumdicker et al. 2012; Collins and Higgs 2012).

Given that the accessory fraction of the pangenome is enriched for loci involved in properties from toxin production, to restriction-modification systems, and surface antigens, to say nothing of drug-resistant genes, it is hard to imagine that it might fit well to a neutral model—in several cases though, it does (Baumdicker et al. 2012; Marttinen and Hanage 2017). This result is hard to accept, and it should be given all we know about the power of selection and the size of bacterial populations. However, it should be appreciated that a multitude of selective scenarios can produce a signal that is hard or impossible to distinguish from neutrality. Study of other metrics may be required to unveil the underlying processes. For instance, the rates with which diverging strains of pneumococcus acquired or lost genes was found to be indistinguishable from neutrality and even to yield good estimates of the population mutation and recombination rates (Marttinen and Hanage 2017). Yet later analysis of the same population, alongside others, was interpreted as strong evidence for negative frequency-dependent selection on the accessory fraction of the genome (see above). What is going on?

A possible explanation lies in the central limit theorem. If an outcome is determined by many independent random variables, each with finite variance, then we expect the result of adding them all together to be a normal distribution. In other words, if the fitness of a strain is the consequence of many independent factors, we might find it appears neutral—the chances of any individual getting into the next generation could be normally distributed around 50:50. This result has been the source of substantial interest in ecology, given that it can be used to show that species abundance distributions (SADs—a common metric for summarizing ecological diversity (McGill et al. 2007) can appear neutral while actually being the result of many non-neutral processes. In the case of bacteria, the fitness effects of genes on the same mobile element may not be independent, however, the effects of multiple mobile elements may similarly approximate to an overall strain fitness not distinguishable from neutrality. Other models from community ecology may be useful in determining the contents of genomes, as well as ecosystems.

Nevertheless, the current consensus in the field is that gene variation directly reflects the ecological niche occupied by the bacteria (Sheppard et al. 2018) and the response to local selective pressures (Cordero and Polz 2014). This may involve the acquisition of antibiotic-resistant genes, as described above, metabolic genes needed to exploit a novel energy source, bacteriocins for microbial warfare, or phage and phage-defense genes involved in predation–prey “paper-rock-scissor” dynamics, as so eloquently described by Corander and colleagues (Arnold et al. 2018). Further, it is suspected that rapid acquisition and dissemination of genes most often occurs as bacterial clones adapt to a novel niche previously occupied by another species (Polz et al. 2013; Popa et al. 2011; Smillie et al. 2011; Vos et al. 2015). An example of this would be the acquisition of IncA/C plasmid by *Vibrio cholerae* introduced to Haiti, a country previously devoid of epidemic cholera for at least 100 years (Carraro et al. 2016) as well as the post-vaccine population of *S. pneumoniae* in the USA, which experienced a significant population shift after the 7-valent pneumococcal conjugate vaccine removed approximately 30% of the pre-vaccine population. Niches themselves are not explicitly segregated, and therefore one does not have to be vacated to

then be exploited by a newcomer. Gene flow may occur between sympatric lineages; i.e., habitat borders are not defined by walls or other barriers, and recombination can occur among lineages of a species where habitat space is not clearly demarcated (Martinen and Hanage 2017). This model explains lineage divergence and population structure among several species, and is important because it highlights that a species requires not only the ability to acquire genes but also the opportunity to do so. Interestingly, it has also been suggested that once a competent species encounters a new niche, it can give rise to noncompetent lineages, providing an advantage when adaptation through gene acquisition is not required and may, in fact, be deleterious (Jorth and Whiteley 2012).

The acquisition of genes is not always beneficial and may, in fact, be deleterious (Vos et al. 2015). Indeed, for every successful lineage that is observed, there are likely several “failed” ecological experiments. Since there is not a clear delineation between the fitness gain and costs of gene acquisition, it may be an oversimplification of the dynamics to ascribe a net-positive or net-negative effect of gene gain and loss. The truth, of course, is somewhere in between and likely varies to different degrees between species. To offset fitness costs and compensate for the acquisition of mobile elements, mutations may arise in core loci and form epistatic relationships with the acquired gene. This has been suggested, for example, in *E. coli*, where nucleotide substitutions in regulatory genes were found to be associated with the acquisition and maintenance of accessory genes (McNally et al. 2016). This dynamic is further supported, in part, by recent findings of epistatic interactions across genome-wide loci among multiple bacterial species (Arnold et al. 2018; Skwark et al. 2017).

There are examples where ecological niches are clearly defined among species and others where the relationships between habitat and organism are obscured. In the *E. coli* study (McNally et al. 2016), ecological adaptation and niche segregation were not observed among isolates collected from humans and animals, while in other species such as *Campylobacter*, this is commonly observed (Sheppard et al. 2011). Methods to investigate gene flow and selection in the context of adaptation and ecology are continually being refined. In some instances, identifying the appropriate system to test ecological hypotheses is the limiting step. An intriguing approach to understanding these associations is not to identify niches, the organisms that inhabit them, and then attempt to resolve the genes associated with adaptation, but instead first assess gene flow and then make predictions about ecology. So-called “reverse-ecology” proposed by Shapiro and Polz seeks to investigate habitat specificity by assessing gene flow and gene-specific sweeps, and has been used to predict ecological differentiation of *Vibrio* spp. in aquatic environments (Hunt et al. 2008; Shapiro et al. 2012; Shapiro and Polz 2014). They demonstrate an example of applying a fresh perspective to an appropriate model system to understanding bacterial ecology.

Taken together, the accumulation of population samples that have been analyzed with modern genomic methods has greatly improved our understanding of the pangenome, and its ecological significance. The totality of loci in a sample includes the essential core, together with a set of accessory loci that have a range of ecological and evolutionary significance: from functional genes with direct relevance to niche

such as those described in the reverse ecology approach of Shapiro and Polz, to more selfish elements such as toxin–antitoxin systems. One feature of the current landscape of bacterial genomics that is not often noticed, is that for all the references in the literature to “Whole Genome Sequencing,” few studies actually determine the whole, i.e., finished genome including all plasmids. Our current understanding is overwhelmingly based on high-quality draft, not finished, genomes. The emergence of long-read technologies is changing this, and as they improve and become more economical (together with more methods for making hybrid assemblies from short- and long-read data) we may find that our current understanding underestimates the actual quantities of sequence variation in bacteria and that there are short regions under strong selection that accumulate rapid change and are hence hard to assemble from short-read data. Adding these is just one of the exciting directions for research over the next few years, which is sure to improve our understanding of pangenomes and their significance far beyond our current knowledge.

Acknowledgments WPH would like to acknowledge support from U54GM088558 from NIGMS, and helpful discussions with Franz Baumdicker and Claire-Marie MacLean.

References

- Arnold BJ, Gutmann MU, Grad YH, Sheppard SK, Corander J, Lipsitch M et al (2018) Weak epistasis may drive adaptation in recombining bacteria. *Genetics*. <https://doi.org/10.1534/genetics.117.300662>
- Azarian T, Grant L, Georgieva M, Hammitt L, Reid R, Bentley S et al (2016) Pneumococcal protein antigen serology varies with age and may predict antigenic profile of colonizing isolates. *J Infect Dis*. <https://doi.org/10.1093/infdis/jiw628>
- Azarian T, Grant LR, Arnold BJ, Hammitt LL, Reid R, Santosham M et al (2018) The impact of serotype-specific vaccination on phylodynamic parameters of *Streptococcus pneumoniae* and the pneumococcal pan-genome. *PLoS Pathog* 14. <https://doi.org/10.1371/journal.ppat.1006966>
- Baumdicker F (2015) The site frequency spectrum of dispensable genes. *Theor Popul Biol* 100:13–25. <https://doi.org/10.1016/j.tpb.2014.12.001>
- Baumdicker F, Hess WR, Pfaffelhuber P (2012) The infinitely many genes model for the distributed genome of bacteria. In: *Genome biology and evolution*. Oxford University Press, Oxford, pp 443–456. <https://doi.org/10.1093/gbe/evs016>
- Bentley SD, Aanensen DM, Mavroidi A, Saunders D, Rabinowitsch E, Collins M et al (2006) Genetic analysis of the capsular biosynthetic locus from all 90 pneumococcal serotypes. *PLoS Genet* 2:e31. <https://doi.org/10.1371/journal.pgen.0020031>
- Bilek N, Ison CA, Spratt BG (2009) Relative contributions of recombination and mutation to the diversification of the opa gene repertoire of *Neisseria gonorrhoeae*. *J Bacteriol* 191:1878–1890. <https://doi.org/10.1128/JB.01518-08>
- Blanquart F, Lehtinen S, Lipsitch M, Fraser C (2018) The evolution of antibiotic resistance in a structured host population. *J R Soc Interface* 15:20180040. <https://doi.org/10.1098/rsif.2018.0040>
- Carraro N, Rivard N, Ceccarelli D, Colwell RR, Burrus V (2016) IncA/C conjugative plasmids mobilize a new family of multidrug resistance islands in clinical *Vibrio cholerae* Non-O1/Non-O139 isolates from Haiti. *MBio* 7. <https://doi.org/10.1128/mBio.00509-16>

- Chen J, Quiles-Puchalt N, Chiang YN, Bacigalupe R, Fillol-Salom A, Chee MSJ et al (2018) Genome hypermobility by lateral transduction. *Science* 362:207–212. <https://doi.org/10.1126/science.aat5867>
- Cobey S, Baskerville EB, Colijn C, Hanage W, Fraser C, Lipsitch M (2017) Host population structure and treatment frequency maintain balancing selection on drug resistance. *J R Soc Interface* 14. <https://doi.org/10.1098/rsif.2017.0295>
- Collins RE, Higgs PG (2012) Testing the infinitely many genes model for the evolution of the bacterial core genome and pangenome. *Mol Biol Evol* 29:3413–3425. <https://doi.org/10.1093/molbev/mss163>
- Corander J, Fraser C, Gutmann MU, Arnold B, Hanage WP, Bentley SD et al (2017) Frequency-dependent selection in vaccine-associated pneumococcal population dynamics. *Nat Ecol Evol* 1:1950–1960. <https://doi.org/10.1038/s41559-017-0337-x>
- Cordero OX, Polz MF (2014) Explaining microbial genomic diversity in light of evolutionary ecology. *Nat Rev Microbiol* 12:263–273. <https://doi.org/10.1038/nrmicro3218>
- Croucher NJ, Harris SR, Barquist L, Parkhill J, Bentley SD (2012) A high-resolution view of genome-wide pneumococcal transformation. *PLoS Pathog* 8:e1002745. <https://doi.org/10.1371/journal.ppat.1002745>
- Croucher NJ, Finkelstein JA, Pelton SI, Mitchell PK, Lee GM, Parkhill J et al (2013) Population genomics of post-vaccine changes in pneumococcal epidemiology. *Nat Genet* 45:656–663. <https://doi.org/10.1038/ng.2625>
- Croucher NJ, Coupland PG, Stevenson AE, Callendrello A, Bentley SD, Hanage WP (2014) Diversification of bacterial genome content through distinct mechanisms over different time-scales. *Nat Commun* 5:1–12. <https://doi.org/10.1038/ncomms6471>
- Croucher NJ, Mostowy R, Wymant C, Turner P, Bentley SD, Fraser C (2016) Horizontal DNA transfer mechanisms of bacteria as weapons of intragenomic conflict. *PLoS Biol* 14:e1002394. <https://doi.org/10.1371/journal.pbio.1002394>
- Donati C, Hiller NL, Tettelin H, Muzzi A, Croucher NJ, Angiuoli SV et al (2010) Structure and dynamics of the pan-genome of *Streptococcus pneumoniae* and closely related species. *Genome Biol* 11:R107. <https://doi.org/10.1186/gb-2010-11-10-r107>
- Feng Y, Chen C-J, Su L-H, Hu S, Yu J, Chiu C-H (2008) Evolution and pathogenesis of *Staphylococcus aureus*: lessons learned from genotyping and comparative genomics. *FEMS Microbiol Rev* 32:23–37. <https://doi.org/10.1111/j.1574-6976.2007.00086.x>
- Fortier L-C, Sekulovic O (2013) Importance of prophages to evolution and virulence of bacterial pathogens. *Virulence* 4:354–365. <https://doi.org/10.4161/viru.24498>
- Georgieva M, Kagedan L, Lu Y-J, Thompson CM, Lipsitch M (2018) Antigenic variation in *Streptococcus pneumoniae* PspC promotes immune escape in the presence of variant-specific immunity. *MBio* 9:e00264–e00218
- Griffith F (1928) The significance of pneumococcal types. *J Hyg (Lond)* 27:113–159
- Hollingshead SK, Becker R, Briles DE (2000) Diversity of PspA: mosaic genes and evidence for past recombination in *Streptococcus pneumoniae*. *Infect Immun* 68:5889–5900. <https://doi.org/10.1128/IAI.68.10.5889-5900.2000>
- Hunt DE, David LA, Gevers D, Preheim SP, Alm EJ, Polz MF (2008) Resource partitioning and sympatric differentiation among closely related bacterioplankton. *Science* 320:1081–1085. <https://doi.org/10.1126/science.1157890>
- Iannelli F, Oggioni MR, Pozzi G (2002) Allelic variation in the highly polymorphic locus pspC of *Streptococcus pneumoniae*. *Gene* 284:63–71. [https://doi.org/10.1016/S0378-1119\(01\)00896-4](https://doi.org/10.1016/S0378-1119(01)00896-4)
- International Working Group on the Classification of Staphylococcal Cassette Chromosome Elements (IWG-SCC) (2009) Classification of staphylococcal cassette chromosome mec (SCCmec): guidelines for reporting novel SCCmec elements. *Antimicrob Agents Chemother* 53:4961–4967. <https://doi.org/10.1128/AAC.00579-09>
- Jorth P, Whiteley M (2012) An evolutionary link between natural transformation and CRISPR adaptive immunity. *MBio* 3. <https://doi.org/10.1128/mBio.00309-12>

- Kislyuk AO, Haegeman B, Bergman NH, Weitz JS (2011) Genomic fluidity: an integrative view of gene diversity within microbial populations. *BMC Genomics* 12:32. <https://doi.org/10.1186/1471-2164-12-32>
- Lehtinen S, Blanquart F, Croucher NJ, Turner P, Lipsitch M, Fraser C (2017) Evolution of antibiotic resistance is linked to any genetic mechanism affecting bacterial duration of carriage. *Proc Natl Acad Sci USA* 114:1075–1080. <https://doi.org/10.1073/pnas.1617849114>
- Marttinen P, Hanage WP (2017) Speciation trajectories in recombining bacterial species. *PLoS Comput Biol* 13:e1005640. <https://doi.org/10.1371/journal.pcbi.1005640>
- Marttinen P, Croucher NJ, Gutmann MU, Corander J, Hanage WP (2015) Recombination produces coherent bacterial species clusters in both core and accessory genomes. *Microb Genomics* 1. <https://doi.org/10.1099/mgen.0.000038>
- McGill BJ, Etienne RS, Gray JS, Alonso D, Anderson MJ, Benecha HK et al (2007) Species abundance distributions: moving beyond single prediction theories to integration within an ecological framework. *Ecol Lett* 10:995–1015. <https://doi.org/10.1111/j.1461-0248.2007.01094.x>
- McInerney JO (2017) Horizontal gene transfer is less frequent in eukaryotes than prokaryotes but can be important (retrospective on DOI 10.1002/bies.201300095). *Bioessays* 39:1700002. <https://doi.org/10.1002/bies.201700002>
- McInerney JO, McNally A, O'Connell MJ (2017) Why prokaryotes have pangenomes. *Nat Microbiol* 2:17040. <https://doi.org/10.1038/nmicrobiol.2017.40>
- McNally A, Oren Y, Kelly D, Pascoe B, Dunn S, Sreecharan T et al (2016) 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. <https://doi.org/10.1371/journal.pgen.1006280>
- Ochman H, Davalos LM (2006) The nature and dynamics of bacterial genomes. *Science* 311:1730–1733. <https://doi.org/10.1126/science.1119966>
- Polz MF, Alm EJ, Hanage WP (2013) Horizontal gene transfer and the evolution of bacterial and archaeal population structure. *Trends Genet* 29:170–175. <https://doi.org/10.1016/j.tig.2012.12.006>
- Popa O, Hazkani-Covo E, Landan G, Martin W, Dagan T (2011) Directed networks reveal genomic barriers and DNA repair bypasses to lateral gene transfer among prokaryotes. *Genome Res* 21:599–609. <https://doi.org/10.1101/gr.115592.110>
- Reyes-Robles T, Dillard RS, Cairns LS, Silva-Valenzuela CA, Housman M, Ali A et al (2018) *Vibrio cholerae* outer membrane vesicles inhibit bacteriophage infection. *J Bacteriol* 200. <https://doi.org/10.1128/JB.00792-17>
- Rocha EPC, Smith JM, Hurst LD, Holden MTG, Cooper JE, Smith NH et al (2006) Comparisons of dN/dS are time dependent for closely related bacterial genomes. *J Theor Biol* 239:226–235. <https://doi.org/10.1016/j.jtbi.2005.08.037>
- Romero P, Croucher NJ, Hiller NL, Hu FZ, Ehrlich GD, Bentley SD et al (2009) Comparative genomic analysis of ten *Streptococcus pneumoniae* temperate bacteriophages. *J Bacteriol* 191:4854–4862. <https://doi.org/10.1128/JB.01272-08>
- Schönknecht G, Weber APM, Lercher MJ (2014) Horizontal gene acquisitions by eukaryotes as drivers of adaptive evolution. *Bioessays* 36:9–20. <https://doi.org/10.1002/bies.201300095>
- Seed KD, Bodi KL, Kropinski AM, Ackermann H-W, Calderwood SB, Qadri F et al (2011) Evidence of a dominant lineage of *Vibrio cholerae*-specific lytic bacteriophages shed by cholera patients over a 10-year period in Dhaka, Bangladesh. *MBio* 2:e00334–e00310. <https://doi.org/10.1128/mBio.00334-10>
- Shapiro BJ, Polz MF (2014) Ordering microbial diversity into ecologically and genetically cohesive units. *Trends Microbiol* 22:235–247. <https://doi.org/10.1016/j.tim.2014.02.006>
- Shapiro BJ, Friedman J, Cordero OX, Preheim SP, Timberlake SC, Szabó G et al (2012) Population genomics of early events in the ecological differentiation of bacteria. *Science* 336:48–51. <https://doi.org/10.1126/science.1218198>

- Sheppard SK, Colles FM, McCarthy ND, Strachan NJC, Ogden ID, Forbes KJ et al (2011) Niche segregation and genetic structure of campylobacter jejuni populations from wild and agricultural host species. *Mol Ecol* 20:3484–3490. <https://doi.org/10.1111/j.1365-294X.2011.05179.x>
- Sheppard SK, Guttman DS, Fitzgerald JR (2018) Population genomics of bacterial host adaptation. *Nat Rev Genet* 19:1–17. <https://doi.org/10.1038/s41576-018-0032-z>
- Shintani M, Sanchez ZK, Kimbara K (2015) Genomics of microbial plasmids: classification and identification based on replication and transfer systems and host taxonomy. *Front Microbiol* 6:242. <https://doi.org/10.3389/fmicb.2015.00242>
- Skwark MJ, Croucher NJ, Puranen S, Chewapreecha C, Pesonen M, Xu YY et al (2017) Interacting networks of resistance, virulence and core machinery genes identified by genome-wide epistasis analysis. *PLoS Genet* 13:e1006508. <https://doi.org/10.1371/journal.pgen.1006508>
- Smillie CS, Smith MB, Friedman J, Cordero OX, David LA, Alm EJ (2011) Ecology drives a global network of gene exchange connecting the human microbiome. *Nature* 480:241–244. <https://doi.org/10.1038/nature10571>
- Tettelin H, Massignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL et al (2005) Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial “pan-genome”. *Proc Natl Acad Sci* 102:13950–13955. <https://doi.org/10.1073/pnas.0506758102>
- Vos M, Hesselman MC, te Beek TA, van Passel MWJ, Eyre-Walker A (2015) Rates of lateral gene transfer in prokaryotes: high but why? *Trends Microbiol* 23:598–605. <https://doi.org/10.1016/j.tim.2015.07.006>
- Waldor MK, Friedman DI (2005) Phage regulatory circuits and virulence gene expression. *Curr Opin Microbiol* 8:459–465. <https://doi.org/10.1016/J.MIB.2005.06.001>
- Welch RA, Burland V, Plunkett G, Redford P, Roesch P, Rasko D et al (2002) Extensive mosaic structure revealed by the complete genome sequence of uropathogenic *Escherichia coli*. *Proc Natl Acad Sci USA* 99:17020–17024. <https://doi.org/10.1073/pnas.252529799>
- Wiedenbeck J, Cohan FM (2011) Origins of bacterial diversity through horizontal genetic transfer and adaptation to new ecological niches. *FEMS Microbiol Rev* 35:957–976. <https://doi.org/10.1111/j.1574-6976.2011.00292.x>
- Wozniak RAF, Waldor MK (2009) A toxin-antitoxin system promotes the maintenance of an integrative conjugative element. *PLoS Genet* 5:e1000439. <https://doi.org/10.1371/journal.pgen.1000439>
- Wozniak RAF, Fouts DE, Spagnoletti M, Colombo MM, Ceccarelli D, Garriss G et al (2009) Comparative ICE genomics: insights into the evolution of the SXT/R391 family of ICEs. *PLoS Genet* 5:e1000786. <https://doi.org/10.1371/journal.pgen.1000786>
- Young JPW (2016) Bacteria are smartphones and mobile genes are apps. *Trends Microbiol* 24:931–932. <https://doi.org/10.1016/j.tim.2016.09.002>

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the chapter’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.