Chapter 4 from:

Microbial Ecology

Current Advances from Genomics, Metagenomics and Other Omics

https://doi.org/10.21775/9781912530021

Edited by Diana E. Marco

Faculty of Biological Sciences
Córdoba National University
Argentina

and

CONICET
Córdoba
Argentina

 Caister Academic Press https://www.caister.com

Insular Microbiogeography: Three Pathogens as Exemplars

4

James H. Kaufman[1]*, Christopher A. Elkins[2,3], Matthew Davis[1], Allison M. Weis[4], Bihua C. Huang[4], Mark K. Mammel[2], Isha R. Patel[2], Kristen L. Beck[1], Stefan Edlund[1], David Chambliss[1], Judith Douglas[1], Simone Bianco[1], Mark Kunitomi[1] and Bart C. Weimer[4]*

[1]IBM Almaden Research Center, San Jose, CA, USA.
[2]Division of Molecular Biology, Center for Food Safety and Applied Nutrition, United States Food and Drug Administration, Laurel, MD, USA.
[3]Division of Healthcare Quality Promotion, National Center for Emerging and Zoonotic Infectious Diseases, CDC, Atlanta, GA, USA.
[4]University of California Davis, School of Veterinary Medicine, 100K Pathogen Genome Project, Davis, CA, USA.

*Correspondence: jhkauf@us.ibm.com; bcweimer@ucdavis.edu

https://doi.org/10.21775/9781912530021.04

Abstract

Traditional taxonomy in biology assumes that life is organized in a simple tree. Attempts to classify microorganisms in this way in the genomics era led microbiologists to look for finite sets of 'core' genes that uniquely group taxa as clades in the tree. However, the diversity revealed by large-scale whole genome sequencing is calling into question the long-held model of a hierarchical tree of life, which leads to questioning of the definition of a species. Large-scale studies of microbial genome diversity reveal that the cumulative number of *new* genes discovered increases with the number of genomes studied as a power law and subsequently leads to the lack of evidence for a unique core genome within closely related organisms. Sampling 'enough' new genomes leads to the discovery of a replacement or alternative to any gene. This power law behaviour points to an underlying self-organizing critical process that may be guided by mutation and niche selection. Microbes in any particular niche exist within a local web of organism interdependence known as the microbiome. The same mechanism that underpins the macro-ecological scaling first observed by MacArthur and Wilson also applies to microbial communities. Recent metagenomic studies of a food microbiome demonstrate the diverse distribution of community members, but also genotypes for a single species within a more complex community. Collectively, these results suggest that traditional taxonomic classification of bacteria could be replaced with a quasispecies model. This model is commonly accepted in virology and better describes the

diversity and dynamic exchange of genes that also hold true for bacteria. This model will enable microbiologists to conduct population-scale studies to describe microbial behaviour, as opposed to a single isolate as a representative.

Introduction

For over 280 years biologists have classified living organisms using a system first proposed by Carl Linnaeus. This system developed as scientists grouped organisms based on observable characteristics (phenotypes), usually biochemical reactions, into a hierarchical taxonomic tree. This model has guided research ever since (Ruggiero *et al.*, 2015). Today, for the first time, microbiologists are engaged in describing the diversity of microbial life on earth utilizing technological advances in high-throughput sequencing and whole genome sequences (WGS). To obtain a bacterial genome, microbiologists use pure culture techniques to grow and isolate colonies that were commonly called 'clonal', prior to WGS, because the handful of selected biochemical tests were identical among the isolates. Extracting DNA from these colonies, high-throughput sequencing is used to obtain a consensus or average genome (Draper *et al.*, 2017) that is often described to be the prototypical genome. With a conventional world view of taxonomy, this 'clonal' genome is considered to define the organism's taxonomic class (i.e. species) and is thought to have a core or conserved set of genes that can be used to ascribe genes (i.e. traits) for use as unique identifiers that provide a basis for creating a group or clade on a hierarchal tree.

As WGS provides more data, it has become clear that isolated colonies are not clonal, and outbreaks are not represented by clonal organisms; rather, outbreaks are associated with a distribution of pathogenic genotypes. The observed diversity calls into question the concept of a species if enough diversity is captured from various sources of that specific organism. It is possible that the genome diversity is so high that the concept of a prototypical genome that represents a species or even serotype cannot be associated with a single conserved genome. As ever more bacterial genomes are sequenced, more variants are discovered. The taxonomy designed to classify them continues to fragment into ever more classes if the current hierarchal tree approach is maintained.

This diversity has driven the use of WGS to produce reference sequences for use in outbreak identification and investigation. A benefit to this work is production of vast numbers of data accessible for scientific research and comparison that are now uncovering evidence of genome diversity on a scale we have not contemplated. There has been an unprecedented increase in WGS and public release of new genomes is approaching 100,000 per year – a scale that was not anticipated 5 years ago. The repositories are often public and include the National Center for Biotechnology Information (NCBI), Sequence Read Archive (SRA), European Nucleotide Archive (ENA), DNA Data Bank of Japan (DDBJ), Genomic Encyclopedia of Bacteria and Archaea (GEBA), and multinational 100K Pathogen Genome Project (Allard, 2015; Weimer, 2012, 2107; Wu *et al.*, 2009). The crowd-sourced data in these WGS databases has resulted in cataloging of a plethora of microbial genomes and has assisted in the examination of the metagenomics of ecological niches important in medicine, agriculture, energy, and the built environment (Alivisatos *et al.*, 2015; Locey and Lennon, 2016; Shapiro *et al.*, 2012), all of which relies on reference genome databases.

The implementation of large-scale genomics in microbiology gives rise to observations previously impossible from the observations of a few isolates. In particular, the observed

genome diversity challenges long-held beliefs about genome stability, evolutionary rate, and even the definition of a species (Doolittle, 1999; Doolittle and Brunet, 2016). Indeed, the number of unique genes that are represented within a species, known as the 'pan-genome', continues to increase as the expected number of genes conserved across all the members of a species, known as the 'core-genome', diminishes. This powers a new ability to define the trajectory of novel gene discovery to evoke consideration of the advances made by naturalists and applied to animals and theories of animal ecology. With the scale of genomes available the same considerations can be made when examining the diversity of microbial life and microbiome membership. Application of ecological theory provides a new perspective on microbiome structure and enables new insights into community association and membership that have not been examined in detail.

On a macroscopic scale, MacArthur and Wilson's macro-ecological 'theory of island biogeography' is a well-established explanation for species diversity across geographically disconnected groups (Diamond, 1984; Lovejoy *et al.*, 1984, 1986; MacArthur and Wilson, 1963). This theory relates the number of species (S) found, in steady state, within the area (A) of an isolated ecosystem as a power law. Over five orders of area magnitude, S varies as A^z with z, the MacArthur–Wilson exponent, typically in the range of $0.2 < z < 0.3$, decreasing slightly for island groups nearer to continental land masses that enable a slow exchange of species.

MacArthur-Wilson attributes the steady state species–area relationship to the rate of successful immigration of new species (from distant continental reservoirs) and the rate of species extinction, both dependent on the availability of viable niches (MacArthur and Wilson, 1963). The number of niches depends, in turn, on the available land area. Wilson's work also demonstrates that the surviving organisms can, themselves, provide niches for other organisms in the community, including new migrants (Wilson, 1999). As such, the observed scale-free behaviour in macro-ecology depends not only on the geographical distribution of land (Mandelbrot, 1983), but also on the dynamic evolution of the food web (Wilson, 1999). Indeed, ecological models and simulations based on Wilson's work reveal that scale-free behaviour emerges *only* as the food web develops over time (Bak *et al.*, 1988; Kaufman *et al.*, 1998). This process is precisely the mechanism that Bak, Tang, and Wiesenfeld (BTW) predicted would lead to power law scaling in their theory of 'self-organized criticality' (Bak *et al.*, 1988; Mandelbrot, 1983).

Self-organized criticality (SOC) is a ubiquitous mechanism by which complexity arises in nature. The behaviours predicted by the SOC theory have been observed across a wide variety of fields including physics, geology, ecology, and neuroscience (Kaufman *et al.*, 1998; Linkenkaer-Hansen *et al.*, 2001; Smalley *et al.*, 1985). Bak *et al.* (1988) propose that self-organized criticality is a property of any dynamical system that has a critical point *as an attractor*. Such a critical point is often observable experimentally as power law divergence of a correlation length of the system. In physics, critical behaviour often indicates a second-order phase transition. For example, the susceptibility of a ferromagnet diverges as the correlation length of magnetic moments diverges near the critical 'Curie' temperature (Kittel and Holcomb, 1967). Near a critical point, temporal and spatial system variables exhibit power law correlations. The value of the critical exponents for different variables are not the same, but they are not independent either. In fact they are related by simple scaling relations (Tang and Bak, 1988). Dynamical systems exhibit SOC if scaling emerges without the need to tune a control variable or effective temperature. This is the basis for the name of

the theory. The system is self-organized because it tunes itself. The critical point is an attractor of the dynamics. Typically, SOC systems are non-equilibrium but slowly driven (Bak and Paczuski, 1995). Evolutionary systems with low mutation rates satisfy this criterion. Typically, SOC systems have many degrees of freedom such as an ecosystem with a complex web of interdependence.

Metagenomic analysis and microbial life: a large-scale study

In this chapter we describe a large-scale study of gene diversity for thousands of public genomes and three different bacterial genera. Completed in collaboration with our researchers at IBM, the University of California Davis, and the US Food and Drug Administration, this work consists of experiments that quantify the diversity of genes, confirm the power law relationship, redefine taxonomic context, and invoke the quasispecies model for bacteria. To test these principles at the scale of genes and genomes, one must have access to an enormous number of individuals (e.g. WGS) that reflect species diversity. The public databases of WGS provide the underlying data diversity. We took advantage of this large collection to examine the diversity for thousands of public genomes and three different genera to test the applications of BCW theory in the microbial world.

A growing library of genetic data

All WGS data are publicly available via NCBI (Weimer, 2012). The WGS were downloaded, assembled using ABySS (abyss-pe v.1.5.2) (Simpson *et al.*, 2009), and annotated using Prokka (v.1.10) (Seemann, 2014) prior to the analysis. These include new WGS data, sequenced by the 100K Pathogen Genome Project (UC Davis, Davis, CA; Weimer, 2107), as described by Lüdeke *et al.* (2015), using Illumina paired end 100 methods, publicly released on the SRA in the 100K Pathogen Genome Project bioproject (PRJNA186441) and re-assembled for use in this study. To measure the cumulative rate of observation of 'new' genes (both known and 'putative' or unknown) as a function of number of *Salmonella* isolates sampled, 866 individual *Salmonella* isolates were obtained from the 100K Pathogen Genome Project with *de novo* assembly with random subsets of those isolates selected in separate trials, followed by cumulative gene count determined for each trial.

To measure the cumulative rate of observation of 'new' *Campylobacter* genes as a function of the number of isolates sampled, the raw sequence data from 15,158 individual *Campylobacter* isolates from public sources (Leinonen *et al.*, 2010) were assembled and annotated as described for *Salmonella*. The number of genomes represents all available *Campylobacter* WGS data in the SRA, at the time of this publication, for which assembly was successful. The large number was chosen to test the observed scale-free behaviour over four orders of magnitude in the bootstrap as a function of the number of genomes.

Identification of core genes to build the phylogeny for *E. coli* was carried out with the Basic Local Alignment Search Tool (BLAST) with a criterion of 95% identity over 90% of the gene length, querying 42 closed reference genomes. An allowance was made whereby a core gene was retained if missing in no more than one genome. Alignment of the 3348 strain sequences for each gene yielded a single nucleotide polymorphism (SNP) set that was used to construct a phylogeny. From the phylogeny, 334 genotypes were selected to represent the diversity in the tree.

Whole RNA metagenomes were produced using HiSeq 4000 or X instruments with one sample per lane, resulting in ~350 million reads/sample. Genomic distances were determined using the Meier-Kolthoff method (Meier-Kolthoff *et al.*, 2013). Whole genome–genome distance matrices were translated into the Newick tree format using Mega7 with the neighbour-joining method. Genome distance matrixes were clustered and visualized using Matlab and R (Kumar *et al.*, 2016). Genome–genome similarity is defined as 1.0 – genome–genome distance.

Diversity of genes

To quantify genome diversity, we examined the specific gene content by comparing individual proteins from thousands of genomes from three different genera that occupy similar niches in animals: *Salmonella*, *E. coli*, and *Campylobacter*. All genomes were derived from publicly available WGS datasets where each organism was cultured, isolated, and sequenced. We performed a *de novo* assembly using ABySS (abyss-pe v.1.5.2) (Simpson *et al.*, 2009) and gene annotation using Prokka (v.1.10) (Seemann, 2014) for each genome and then quantified the cumulative number of total genes (i.e. proteins) discovered as a function of number of isolates (genomes) sampled in multiple bootstrapped trials (Fig. 4.1).

Each point in Fig. 4.1 represents the cumulative number of distinct amino acid sequences (proteins) inferred for regions annotated as known or hypothetical proteins for one genome, plotted against the number of genomes or isolates in that subset, both known and unknown. The points were derived from 100 separate bootstrapped trials.

For the 866 individual *Salmonella* genomes obtained and re-assembled, the log of the genome diversity grows approximately linearly with the log of the number of isolates (Fig. 4.1a), indicating a power law relationship with the average rate of cumulative gene discovery, N, increasing with the number of isolates, η, as shown in Equation 4.1:

$$N \propto \eta^{0.468 \pm 0.001} \ (Salmonella) \tag{4.1}$$

The same power law (within experimental error) was observed in an analysis of *E. coli* genomes (Fig. 4.1b). In this experiment, 3348 independent genome sequences were selected by classifying them into 334 genotypes to condense the data and search for a common reference representative with < 0.2% genetic difference. As with *Salmonella* (Fig. 4.1a), a power law emerged for *E. coli* relating the average rate of cumulative gene discovery, N, to the number of genotypes, η, as shown in Equation 4.2:

$$N \propto \eta^{0.462 \pm 0.002} \ (E.\ coli) \tag{4.2}$$

To determine whether genome size and/or mutation rate contributes to variation in the exponent, we extended this investigation to another organism, *Campylobacter* (Fig. 4.1c). *Campylobacter* is often found in the same niche (e.g. chicken microbiome) and is known to have a much higher mutation rate than either *E. coli* or *Salmonella* (which have similar mutation rates). To compensate for the additional genetic inclusion rates, we increased the number of genomes to 15,158. This greatly expanded the scale of the genomic calculation yet produced a similar exponent to the other organisms, again demonstrating that genome diversity followed an approximate power law relating the rate of cumulative gene discovery, N, to the number of genotypes, η, as shown in Equation 4.3:

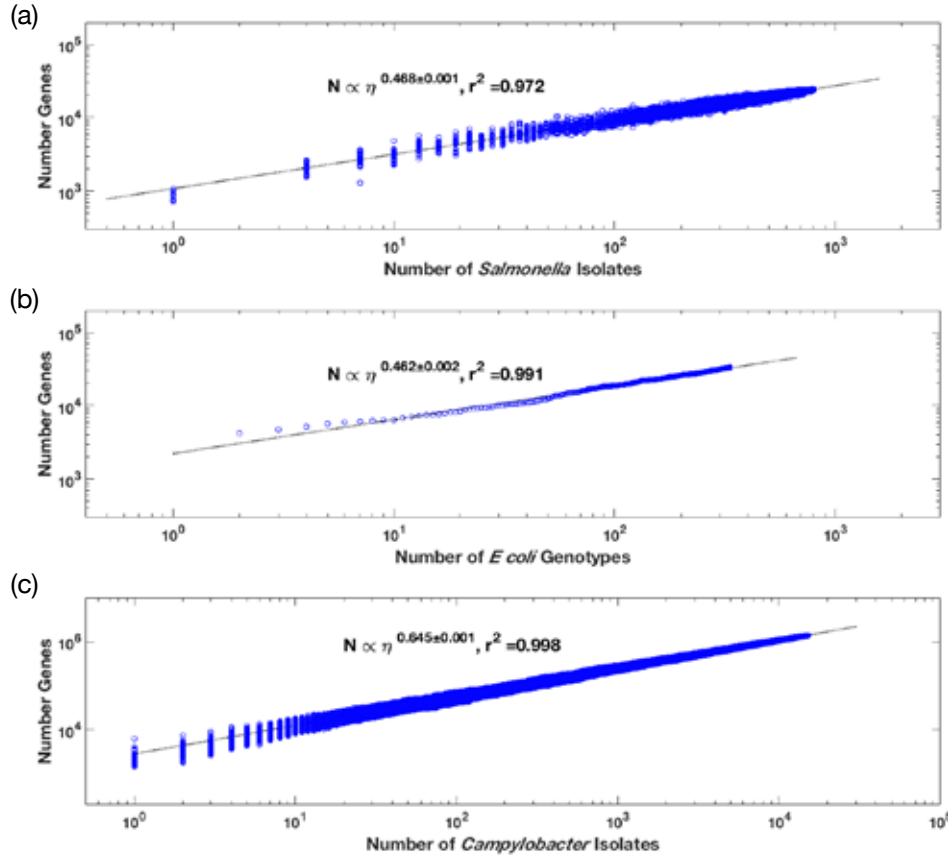

Figure 4.1 The cumulative rate of discovery of 'new genes' as a function of the number of (a) 866 *Salmonella* isolates, (b) *E coli* genotypes from 3348 representing 334 genotypes, and (c) 15,158 *Campylobacter* isolates. All increase as a power law (linear on a log-log plot) defined by Equations 4.1–4.3.

$$N \propto \eta^{0.645 \pm 0.001} \ (Campylobacter) \tag{4.3}$$

In Fig. 4.1, we deliberately use the same symbol (η) to denote both number of isolates (Fig. 4.1a andc) and number of genotypes (Fig. 4.1b). Each genotype represents a collection of isolates grouped in a data reduction effort. The power law behaviour observed in these studies of three different genera at completely different scales suggests a possibly universal functional relationship between the rate of new gene discovery and the number of genotypes or isolates over a wide range in scale. This type of scale-invariant behaviour does not spontaneously emerge in nature, but must reflect an underlying dynamical (evolutionary) process (Bak *et al.*, 1988).

Taxonomic content

As the number of bacterial WGS increases to over 330,000 in public databases, the emerging genotypic diversity is challenging traditional taxonomic hierarchical structures and

concepts of evolution (Al-Saari *et al*., 2015). Classification and naming of organisms with traditional methods is becoming problematic, as the addition of a single new genome drives continual revision of reference phylogenetic trees. For example, recently discovered cryptic environmental lineages of *E. coli* do not fit with current multi-locus sequence typing (MLST) classification for *E. coli sensu stricto* (Walk *et al*., 2009). WGS analysis with a single gene, such as 16s rRNA, provides one name, but whole genome alignment provides a mixture of names where portions of the genome align to more than one genome, demonstrating genome plasticity, widespread homology among gene groups, and possible horizontal gene transfer at a much higher frequency that previously observed.

A similar conclusion was hinted at in a comparative study of lactic acid bacteria that led to the re-alignment of two entire genera (Makarova *et al*., 2006). As the number of WGS data continues to grow at an unprecedented rate, conventional methods to classify organisms using genomics may not meet the standard of absolute identification based on phenotype. Whole genome comparison methods now available circumvent the assumptions currently used to define a constant or static set of single genes, SNPs, or even core sets of genes. Whole genome comparison eliminates the need for data reduction and avoids the associated skewing of taxonomical classification. Although there is a regulatory need to classify and name organisms, the data suggest that classification of bacteria, based on whole genomes, should be considered more akin to the classification of virus in terms of quasispecies, as discussed below. Methods that identify new organisms from metagenome sequence are emerging, making it impossible to use phenotypic characteristics as there is no pure culture (Parks *et al*., 2017). These advances are begging for new methods to classify organisms based on genotype.

To illustrate how whole genome data challenge traditional taxonomy, it is informative to carry out whole genome comparison at different scales. Fig. 4.2 is an analysis of three independent *Campylobacter* datasets.

Principal component analysis (Jolliffe, 2002) based on a whole genome–genome difference matrix reveals similar clusters of *Campylobacter* genomes (Weimer, 2012; Weis, 2016). When using only 90 genomes from a study of primates and crows (Fig. 4.2a), three tight clusters appear in the PCA biplot; this result aligns well with classic phylogenetic classification and nomenclature and is congruent with a small sampling of the genome space based on only a few genomes. Using 218 genomes from the Ensembl database (Fig. 4.2b), the clusters for the three species in Fig. 4.2a enlarge and overlap substantially, demonstrating the increased genome diversity for an increase of approximately two times more genomes in the analysis. With 715 genomes from the NCBI SRA (Fig. 4.2c) with the same principal component analysis, the *Campylobacter* genome diversity becomes even more apparent. Taken together, these data demonstrate a rapid expansion of genotypic diversity within species, and of diversification, that overlaps species in same genome space. The whole genome analysis reveals how a relatively small increase in the number of genomes can rapidly expand genome space and blur the concept of species as well as traditional classification structures.

This expansion of genome space reflects the increasing genome diversity that occurs as more genomes are added to an analysis. Expansion of the diversity is especially evident when samples originate from a large collection of organisms from different geographical locations that may be insular regions of diversity (MacArthur and Wilson, 1963; Weis, 2016). The expansion and overlap in Fig. 4.2 also implies that traditional naming conventions are inaccurate and lead to false identification based on artefacts that emerge when using a small set

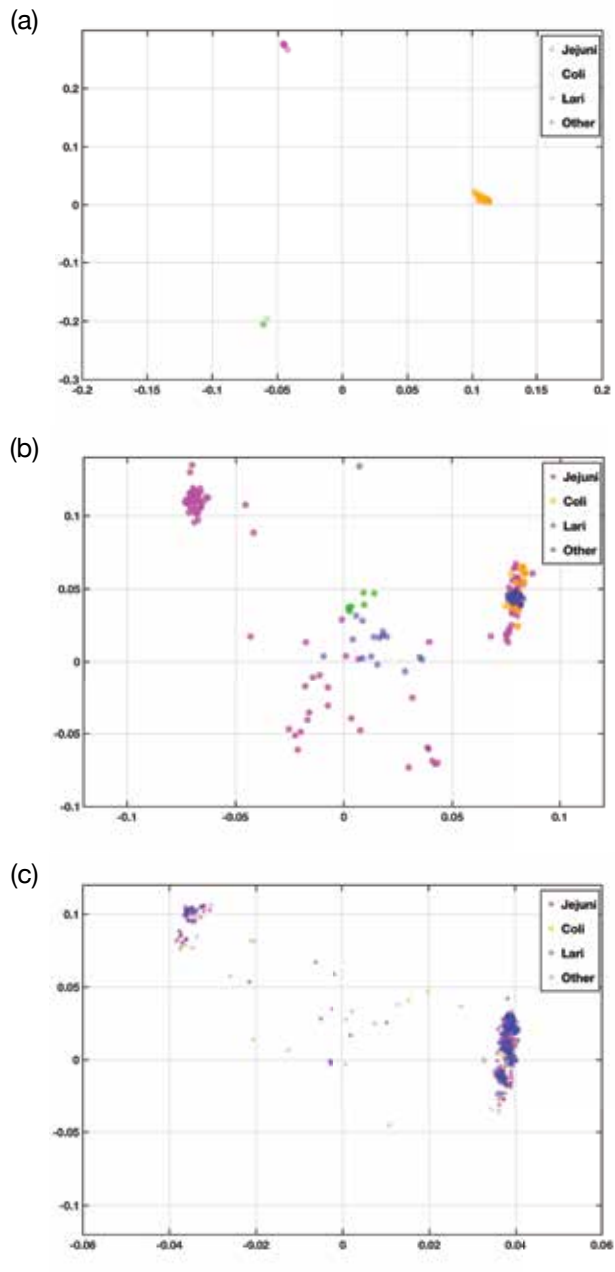

Figure 4.2 Principal component analyses of genome–genome distance from *Campylobacter* genomes representing (a) 90 genomes from primates and crows (Weis, 2016), (b) 218 genomes from the Ensembl reference database, and (c) 715 genomes from the NCBI SRA. In all three figures the *x*-axis is principal component 1 and the *y*-axis is principal component 2. The colour code in the legend indicates the original source species identification. With increasing dataset size, the clades representing traditional views of phylogeny for species begin to expand and overlap.

of input genes, MLST markers, or SNPs. These approaches were useful when computation and data were a limiting factor; however, with more genomes and cloud computing, data reduction methods are not needed and lead to inaccurate or even misleading classification compared with the use of whole genomes and whole genome distance.

Another way to visualize taxonomic grouping is to render a heat map of the full genome–genome distance matrix. The head map is shown in Fig. 4.3, which represents all pairwise distances between *Campylobacter* whole genomes.

With 90 genomes (Fig. 4.3a) the regions of greatest similarity are found along the matrix diagonal. With 218 and 715 genomes (Fig. 4.3b and c, respectively), highly similar but divergent subsets of genomes emerge with increasing numbers of genomes. These appear

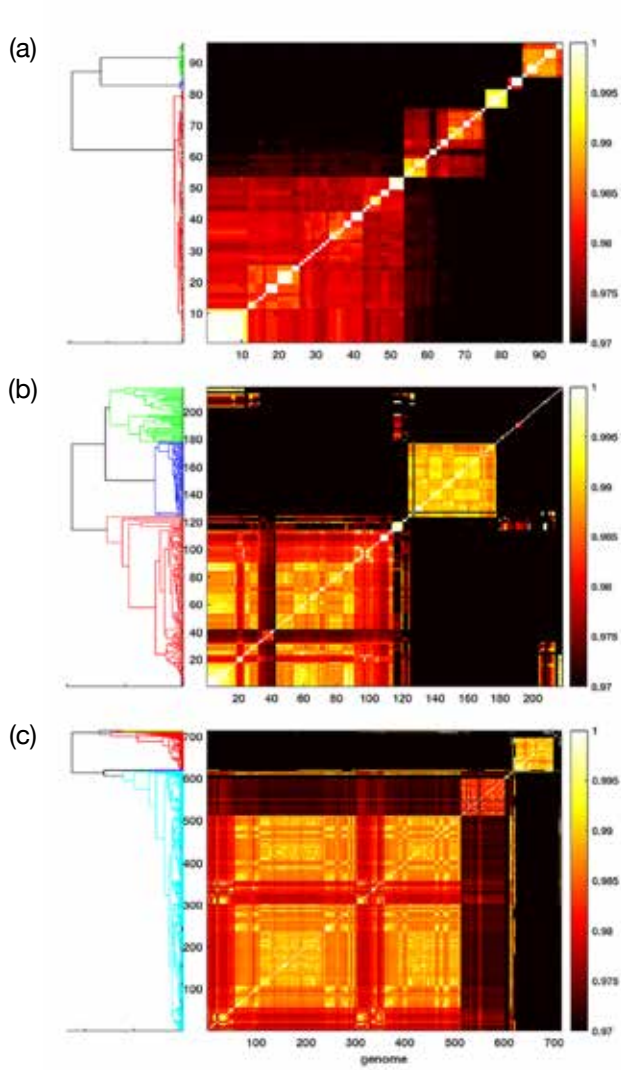

Figure 4.3 Heat maps showing the genome–genome similarity matrix derived from three *Campylobacter* databases containing (a) 90 genomes, (b) 218 genomes, and (c) 715 genomes.

as regions of similarity for well-separated (off-diagonal) genome pairs, indicating similarity between divergent branches of the taxonomic graph. This off-axis similarity reflects processes such as horizontal gene transfer that can result in distantly related organisms sharing similar or identical genes.

When 715 genomes were compared (Fig. 4.3c), the matrix reveals regions of genetic similarity for even more distant taxa, indicating gene transfer between different species.

Analogous behaviour is observed in divergent branches of *E. coli* and *Salmonella enterica* as evidenced by FDA microarray data, first reported by a number of other studies (Elkins *et al.*, 2013; Jackson *et al.*, 2011; Patel *et al.*, 2016).

Whole genome microarray hybridization using 1094 *E. coli* genomes from the FDA *E. coli* Identification (ECID) microarray (Fig. 4.4a) (Patel *et al.*, 2016) and 600 *S. enterica* profiled on a *S. enterica–E. coli* (SEEC) multispecies microarray (Fig. 4.4b) (Elkins *et al.*, 2013) reveal similar findings to the *in silico* analysis.

The majority of *E. coli* and *Salmonella* profiled by this method were amalgamated from clinical, foodborne, and associated environmental strains with relevance to public health and food safety with strains from private, academic, and publicly available collections. The microarray profiling program was developed by the FDA for track-and-trace molecular epidemiology (Elkins *et al.*, 2013). It was born out of an unconventional application towards genomic profiling from its original utility as a gene expression tool. This is primarily because forensic strain-level attribution became possible from the significant intraspecies genome plasticities observed in the resulting profiles (Jackson *et al.*, 2011). In aggregate, whole genome analyses reveal that individual species and strains could differ on the order of megabases using the entire genome. This amount of genomic variation raises questions about how to implement genome-based methods hierarchical classification methods, and points to a quasispecies approach for microbial forensics.

Shapiro *et al.* (2012) attributed this diversity expansion to horizontal gene transfer whereas Doolittle (1999) prophetically described gene sharing beyond what was thought

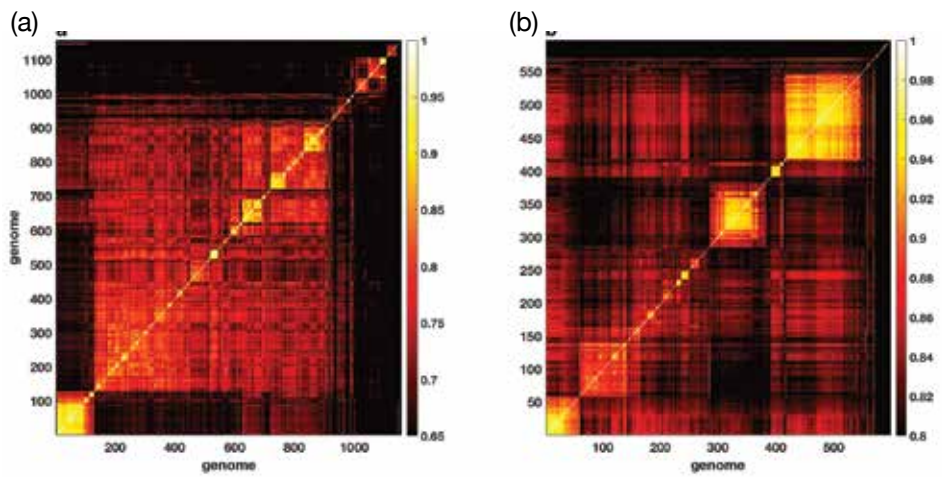

Figure 4.4 Heat maps of genome–genome similarity (Pearson correlations) of pangenome gene (allele) presence measured with custom-designed Affymetrix microarray platforms. (a) Hybridization intensities for 1094 *E. coli* genomes; (b) hybridization intensities for 600 *S. enterica* genomes.

possible. Subsequently, many studies have established that bacteria engage in extensive horizontal gene transfer (Doolittle, 1999; Doolittle and Brunet, 2016; Hug *et al.*, 2016; Woese, 2004). Horizontal gene transfer suggests that 'the tree of life' is not a simple tree (Hug *et al.*, 2016), but rather a complex web of genes that have uneven movement between organisms. The diversity apparent in Figs 4.1–4.4 suggests that bacterial life is organized in a highly interconnected network, a graph containing edges that connect phylogenetic quasispecies across varying genetic distances or scales.

The quasispecies model

Each of the thousands of public genomes analysed was extracted from some microbiome in which organisms existed in a web of organism interdependence. The communities were different, and the organisms and genes (or metagenomes) that existed depended on local conditions. In many cases, the local environment was defined by a macroscopic animal host. As such one might expect scaling behaviour observed in macro-ecologies to extend down to their microbiomes.

Our experiments reveal a power law scaling of the number of new genes versus the number of samples or genomes studied. Mathematically, this relationship rules out the existence of a core genome. As new genomes are studied, variations are eventually found for every gene. Some of these variants may represent slight modifications in genotype (single nonsynonymous mutations). Others represent entirely new phenotypes. We derive a scaling law below that relates the exponent measured here to the MacArthur and Wilson exponent, z. Our analyses of public genomes reveal a common exponent for closely related bacteria (*Salmonella* and *E. coli*) and a slightly larger exponent for a distant organism (*Campylobacter*) with a higher mutation rate and a smaller genome. All values of z are in agreement with the range observed by Wilson and MacArthur. This scaling behaviour reveals diversity that emerges from examining over 15,000 public samples. Experimental analysis of 40 metagenomic samples confirms large-scale genotypic diversity of *Campylobacter* in a single microbiome.

The scaling behaviour empirically observed requires us to rethink the system of classification used to define species. Eigen and Schuster suggest that a 'cloud' of diverse but related organisms within a population are more accurately represented as quasispecies with a distribution of genomes and a distribution of genes (Eigen and Schuster, 1977, 2012). This concept is commonly invoked in virology but data demonstrating its applications to bacterial evolution have been limited (Allard, 2015; Jolley *et al.*, 2004). Available genome sequences are just now reaching a critical mass required to make the observations demonstrated here (Allard, 2015; Leinonen *et al.*, 2010; Tateno *et al.*, 2002; Wu *et al.*, 2009).

The public genomes analysed here were submitted to the SRA by multiple researchers performing a wide variety of independent studies. In these studies, samples were derived using laboratory procedures where each pathogen was studied as a single genotype, reflecting a traditional view of phylogeny where organisms are cultured, isolated, and sequenced from single colonies. In fact, organisms exist in genomically diverse populations or communities of related, but not identical genotypes – even within a colony. These communities evolve rapidly under selective pressure by a number of mechanisms including horizontal gene transfer, methylation, and plasmid content. Culture organisms for sequencing as reference genomes are recognized as the *average genome of a colony*, in contrast to an oversimplified model of clonal outbreaks. In fact, these organisms exist in highly diverse populations of

related genotypes that reproduce with high mutation rates including point mutations, larger scale insertions and deletions, and gene transfer. Shapiro *et al.* (2012) reported ecologically driven differentiation of genes in recently diverged populations of ocean bacteria. Although the number of samples in their study was not large enough to observe scale invariance, it demonstrates the rapidity with which gene transfer can occur across multiple related strains in response to environmental pressures: 'genomic fragments can sweep through populations in an ecology-specific manner … with a clear bias towards within-habitat sharing of DNA' (Shapiro *et al.*, 2012).

The theory of quasispecies postulates an evolutionarily optimized mutation rate, which increases in response to stress. This provides for rapid adaptation including, for example, a very high rate of drug escape. Moreover, high recombination and horizontal gene transfer rates are known to play an important role in increasing both the adaptation rate and genetic diversity. Bacterial populations are known to exhibit high intrinsic mutation and recombination rates, throughout the course of an infection (Suerbaum and Josenhans, 2007) and/or during the acute phase of infection (Linz *et al.*, 2014). Increased mutation rates in bacteria are also apparent during selection pressure, either by natural predators (Weitz *et al.*, 2005) or through expression of alternative error-prone mutator genes (Ebrahimi-Rad *et al.*, 2003). Several *Campylobacter* strains are known to show increased mutation rates (Parkhill *et al.*, 2000), with instances of *C. jejuni* and *C. coli* hypermutator phenotypes linked to the emergence of ciproxin resistance.

The quasispecies model provides a framework to understand fitness and evolvability of a population (Jones *et al.*, 2015; Stern *et al.*, 2014; Xiao *et al.*, 2016). Fitness is determined not by the genetic characteristic of a single, isolated, static species or gene, but by the collective distribution of the members of the quasispecies whose clonal expansion reflects their underlying genotypic diversity and their fitness with respect to a particular environment or selection pressure including antibiotics, other medications, and treatments (Hu and Zhu, 2016; Qin *et al.*, 2010). The model that a particular pathogen associated with an epidemic or outbreak is accurately represented as a single species (e.g. a microbe's name using 16s sequencing) with a clonal identity is not only inadequate but also fallacious: it ignores the underlying genotypic diversity.

Metagenomic techniques make it possible to profile the genetic diversity of a quasispecies within a microbiome. We used RNAseq to study the metatranscriptome of the microbial community of 27 different poultry meal samples. Combining data from all samples provides ≈9.5B reads of raw sequence. Although metagenomics is often used to profile the community ecology, we used the data here to profile the genotypic diversity of one organism in the community, *Campylobacter*. These reads were aligned to the 218 *Campylobacter* genomes from the Ensembl database and the alignments tallied for each genome at 95% identity. This reference database is larger than is typically used to identity a single species in a metagenomic study, but small enough to illustrate two alternative views of what that identification means.

A classic analysis might treat the 218 *average genomes* in the reference database as independent genotypes, or clonal leaf nodes, on the tree of life. Using this model, if any of these genotypes were present in the community of the poultry meal, it would be possible to set a strict threshold and ignore any evidence of occurrence in the sample below that threshold.

An alternative analysis with a different perspective might treat the 218 genomes in the reference database as approximate, historic, average observations of genotypic *distributions*

from 218 ecological niches. Alignments to this reference provide a measure of the probability with which genotypes in the *Campylobacter* quasispecies in the poultry meal are similar to those historic genotypes. Genotypes with several thousands of alignments may reflect minority alleles within the community.

Fig. 4.5 shows a circle plot mapping the log-normalized alignments from our metagenomic study of poultry meal to 218 *Campylobacter* reference genomes.

The log-normalization highlights the genotypic diversity reflected by both the reference database and the sample. In this view, it is still possible within the circle plot to label specific genotypes with alignment scores above a fixed threshold, but displaying the alignments to the full reference database reveals that there is, in fact, a distribution of closely related genotypes within the sample.

As WGS and metagenomics gain acceptance in fields like microbiome medicine, outbreak definition, and food safety traceability, where the demands for accuracy and precision are very high, what really matters is function because it is genes that drive the epidemiology. As shown in Fig. 4.1, hundreds of samples and tens of thousands of individual genes are required to reveal scaling behaviour and/or to estimate a power law exponent. To do this accurately requires inclusive and expansive reference databases. If too few references

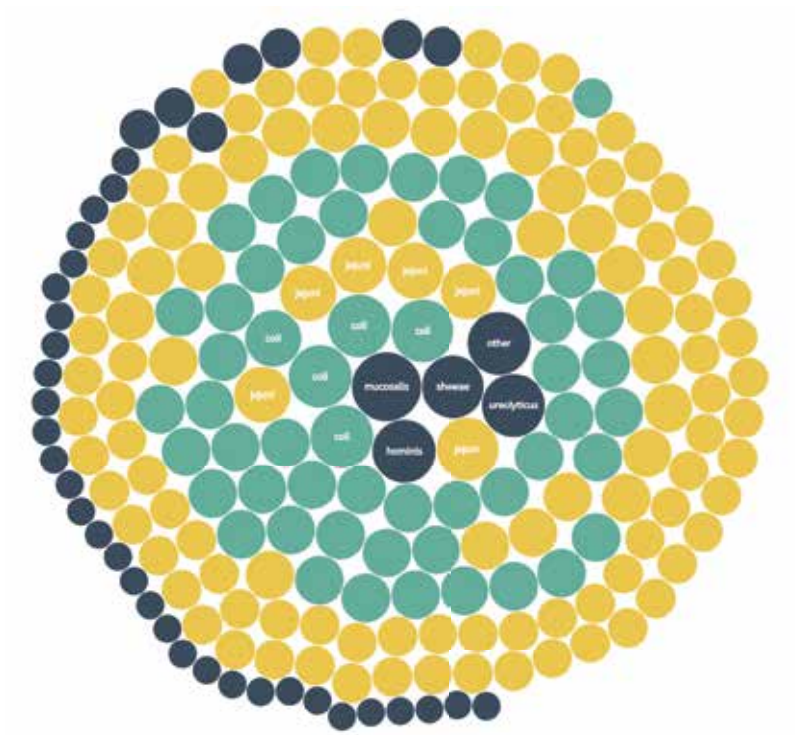

Figure 4.5 A circle plot showing the mapping of all log-normalized read alignments from a metagenomic study of poultry meal to 218 *Campylobacter* genomes in the curated Ensembl database. Each circle represents one specific genome in the reference. The area of each circle is the log of the number of sequence reads that matched that genome. The arrangement of the small circles is arbitrary. The dominant *Campylobacter* species are indicated in the legend.

are in the database, inaccurate associations may be made leading to inappropriate attribution of cause and effect. These alignments reveal the genetic diversity of the *Campylobacter* quasispecies within the chicken meal microbiome based on the 218 curated genomes in Ensembl (Hubbard *et al.*, 2002).

Another approach to measuring diversity is to quantify the allelic variation observed in all reads that align to *Campylobacter*. To do this we enumerate all specific SNP variants and plot the frequency with which these occur (Fig. 4.6).

A particular *Campylobacter* SNP variant (where the nucleotide differs from the reference) that is found only once at a particular location is the most frequent event in the graph. Other variants, where the same substitution is found up to tens of thousands of times at a particular location, are found infrequently. The unique SNP variants were enumerated for all 218 genomes in the reference. The frequency of occurrence for the number of unique SNP variants is approximately a power law with slope near −1 (i.e. $1/f$).

We note that the evidence for scale invariance down to the level of single SNPs does not prove that all nucleotides within a genome are replaced with the same frequency or with the same scale-free behavior. Rather, the highest rate of substitution is typically observed near the 3′-ends of a gene. Substitutions near the 3′-ends are less likely to disrupt function, and more likely to survive (by natural selection).

Insular microbiogeography: an emerging concept for microbiology

The MacArthur–Wilson theory of island biogeography purports that there is an underlying power law relationship between species and habitat on a macroscopic level. Our experimental

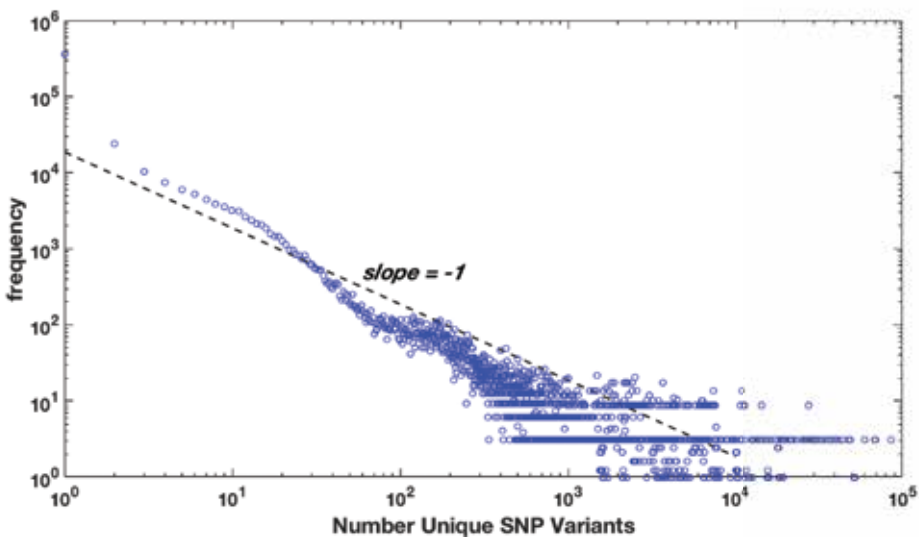

Figure 4.6 The occurrence frequency of unique *Campylobacter* genome SNP variants within a metaRNAseq study of a poultry meal. The function is approximately a power law with slope near −1.

results lead us to posit that the power law exponent that MacArthur and Wilson measured for macroscopic life also applies to microbial life. This is, by definition, scale-free behaviour and leads to some 'laws' that microbes live by to produce genomic variations that we can use as tools for many applications:

- Each host provides a unique environment for community of microorganisms. These microorganisms adapt to their environments.
- Because the distribution of host organisms obeys the MacArthur–Wilson law, so too will the genomic diversity of the microbial communities that colonize them.

This follows from the mathematical definition of scale-free behaviour and from ecological arguments. Microbial communities are linked by the larger web of hierarchical interdependence that links their macro-ecological hosts. To test this proposition, it remains to connect the exponent observed here to the MacArthur–Wilson exponent (Tang and Bak, 1988). The data above show the number of new genes compared with the number of samples or genotypes (Fig. 4.1), whereas MacArthur and Wilson's law relates species or taxa to *geographical area*. Moreover, the data here come primarily from sources originating from human patients and, if the human hosts are interpreted as a single niche, there may seem to be no reason to expect scaling behaviour or any diversity within the genotypic clades. Contrary to this expectation, we find the data indicate a possibly universal power law relationship as a function of number of genomes. In fact, medical sequencing of pathogens is indicated in the event of human disease, typically in unusual cases. As such, these reference sequences do not specifically represent *steady state populations*, but rather originate from consumption of contaminated food or water, or exposure to other disease vectors. Given a global food supply chain and global travel patterns, the simplest interpretation of the data is that these genomes represent a *near random sampling of foodborne pathogens* globally. Humans are engaged in this random sampling as part of everyday consumption. With random sampling, the number of isolates (or genotypes), η, in Fig. 4.3 is related to area, A, as a random walk through sample space. Formally, this implies a functional relationship between area and the number of genomes as $A \approx \eta^{1/2}$. The MacArthur–Wilson exponent, z, is then one-half the exponent obtained in Fig. 4.1 or:

$$N \propto \eta^{2z} \tag{4.4}$$

For all organisms studied here, the data put the corresponding MacArthur–Wilson exponent in the range $0.23 < z < 0.32$, well within the range observed for species-area scaling in the theory of island biogeography.

Equation 4.4 and the data in Fig. 4.1 indicate that as the number of genomes studied increases, the cumulative number of genes observed will continue to increase. The more genomes included in an analysis, the more genes will be discovered (Jacobsen *et al.*, 2011). Although our three experiments involved tens of thousands of genomes, none provided evidence for an upper limit to the rate of gene discovery. Such a limit would manifest as a 'knee' or bend on the log-log plots above some maximum of genomes. For example, in Fig. 4.1c, with 15,158 genomes and 1000 jackknife trials, the study spans a scale exceeding four orders of magnitude with no upper scaling limit found. These observations in three different organisms indicates that a core genome does not likely exist and that as more genomes are

deposited the size of genome space will increase with no practical upper limit – therefore no boundary between species but rather a gradient of change from one to another.

The data in Fig. 4.1a–c reveal that the more genomes we sequence, the more genes will be discovered. For any taxonomic group, at the measured rate of gene discovery, the question becomes, 'How many samples would be required before an alternative gene is discovered for every gene in any randomly selected genome?'

The answer to this question is a special case of the 'coupon collector's problem' (Blom *et al.*, 2012). If we assume that non-synonymous mutations are possible for any gene, then the question becomes, after finding N new genes at random from a genome of length N_o, at what value of N does the probability of *not finding* an alternative gene for every gene in the original genome become vanishingly small.

Let $Z_i^N Z_i^r$ represent the event that the i-th gene is not replaced after N new genes are found. Then, i is r :

$$P\left(Z_i^N\right)=\left(1-\frac{1}{N_o}\right)^N \le e^{-N/N_o} \tag{4.5}$$

where the inequality follows from the Taylor expansion of e^{-x} for small x.

For $N > 2\,N_o \ln(N_o)$:

$$P\left(N > 2\ N_o\ ln\left(N_o\right)\right)=P\left(\cup_i Z_i^{2\ N_o\ ln(N_o)}\right)\le N_o P\left(Z_i^{2N_o\ ln(N_o)}\right)\le\frac{1}{N_o} \tag{4.6}$$

Therefore, the probability of not finding an alternative to *every gene* falls below $1/N_o$ for $N > 2\,N_o \ln(N_o)$. For *Campylobacter*, with $N_o \approx 1500$ genes, this requires finding $N \approx 22{,}000$ new genes. From Fig. 4.1 this will take place after randomly sampling just a few hundred isolates. For *Salmonella*, a variant to each of $N_o \approx 4500$ genes will be found after discovering ≈76,000 genes, which requires approximately 2000 isolates. However, this was not the case and the logical conclusion is that a unique core genome does not exist (DeLong *et al.*, 2010; Green *et al.*, 2004; Horner-Devine *et al.*, 2004).

Implications for microbiology

The evidence for scaling over several orders of magnitude in sample diversity of microbial genotypes and genomes has practical implications for microbiology and the way we catalogue organisms and investigate outbreaks. Approaches that depend on data reduction or definition of a core genome, selection of small numbers of SNPs, or of core MLST gene sets, are destined to fail with false negatives as they do not capture the expansive genome diversity that is unavoidable. The fact that adding one more genotype to the set can indicate a need to rebuild the taxonomy shows that we have not adequately sampled the genomic space and that the tree of life for bacteria is not a tree – it is a network (or graph) of interacting genomes within an environment or geography (Al-Saari *et al.*, 2015; Makarova *et al.*, 2006; Walk *et al.*, 2009), with loops that indicate gene transfer between divergent taxa.

As in virology, it is more appropriate to think of bacteria in terms of quasispecies with a distribution of genotypes and genes that evolve, move, and shift in genome space in response to a changing fitness landscape. The tools required for whole genome comparison using the entire genome exist and must be applied to capture genetic similarity between both nearby and distant genotypes. This is necessary for accurate identification and functional

association required by applications in microbiome-induced physiology, outbreak detection, and food regulation. Adopting whole genome comparison may force microbiology to replace the practice of naming and even the concepts of 'genus' and 'species' with new data models and notations that label populations by their function and the niche(s) they occupy. From a regulatory and health perspective, it is most important to know if an organism has a pathogenic gene that may cause illness (regardless of its taxonomic classification).

There are several dimensions one might consider for a new nomenclature system. New notations could capture the diversity or statistical distributions that describe the frequency with which important function exists within a community. Community level techniques including metagenomics and meta-transcriptomics can be used to predict the probability that specific genes are (or are not) present in a community and the likelihood with which they may emerge as hazards in response to fitness pressures.

Conclusions: gene discovery and power law scaling

The discovery rate of new genes depends on number of genomes sequenced as a power law. If this scaling behaviour arises from the hierarchical web of species interdependence, one should expect the MacArthur–Wilson theory of insular biogeography to apply to microbes. MacArthur–Wilson never derived or explained the power law behaviour they discovered. That explanation comes from the more general theory of self-organized criticality (Bak *et al.*, 1988). The power law reflects the complexity of the underlying food web that connects the macroscopic and microscopic species with their chemical environment. Observation of the same scale-invariant behaviour for microbes and higher order life suggests this food web evolves towards a critical state with fluctuations expected on all scales. To test this proposition, we derive a scaling relation relating the observed exponent to the MacArthur–Wilson exponent, assuming that genome datasets in public repositories represent a near random sampling from the global genome space. The values obtained are consistent with theory. Species require niches to survive and thrive. With respect to microbes these niches are not restricted to the physical and chemical properties of insular environments (which themselves evolve a characteristic fractal geometry) but also include other species. It is the interacting web of ecological dependence between species that leads to self-organization, evolution of a scale-free diversity of life forms across scales within isolated ecosystems. The host organisms are linked to their microbiomes. They provide unique habitats for the communities of microbes that colonize them.

The diversity predicted by a theory of insular microbiogeography requires we rethink our approach to organism classification and regulation. Bacteria form quasispecies related by an slowly evolving phylogenetic graph – not a tree – that reflects the complex web of interdependence spanning all phylogenetic ranks down to the level of microbes. In any open ecosystem or open microbiome, the classical concept of core genomes ceases to be meaningful for any class of bacteria. Classification and regulation based on comparison of whole genomes and *gene function* may better capture the genotypic relationships between related microorganisms.

Acknowledgements

The authors would like to acknowledge contributions from and discussions with Nyuget Kong, Bob Baker, Peter Markwell, Dylan Storey, Barbara Jones and Kenneth L. Clarkson.

References

Al-Saari, N., Gao, F., Rohul, A.A., Sato, K., Sato, K., Mino, S., Suda, W., Oshima, K., Hattori, M., Ohkuma, M., *et al.* (2015). Advanced microbial taxonomy combined with genome-based-approaches reveals that *Vibrio astriarenae* sp. nov., an agarolytic marine bacterium, forms a new clade in Vibrionaceae. PLOS ONE *10*, e0136279. https://doi.org/10.1371/journal.pone.0136279

Alivisatos, A.P., Blaser, M.J., Brodie, E.L., Chun, M., Dangl, J.L., Donohue, T.J., Dorrestein, P.C., Gilbert, J.A., Green, J.L., Jansson, J.K., *et al.* (2015). MICROBIOME. A unified initiative to harness Earth's microbiomes. Science *350*, 507–508. https://doi.org/10.1126/science.aac8480

Allard, M. (2015). GenomeTrakr: A Pathogen Database to Build a Global Genomic Network for Pathogen Traceback and Outbreak Detection. Paper presented at 2015 Annual Meeting (July 25–28, 2015), DesMoines, IA, USA.

Bak, P., and Paczuski, M. (1995). Complexity, contingency, and criticality. Proc. Natl. Acad. Sci. U.S.A. *92*, 6689–6696.

Bak, P., Tang, C., and Wiesenfeld, K. (1988). Self-organized criticality. Phys. Rev. A *38*, 364.

Blom, G., Holst, L., and Sandell, D. (2012). Problems and Snapshots from the World of Probability (Springer Science & Business Media, Berlin).

DeLong, J.P., Okie, J.G., Moses, M.E., Sibly, R.M., and Brown, J.H. (2010). Shifts in metabolic scaling, production, and efficiency across major evolutionary transitions of life. Proc. Natl. Acad. Sci. U.S.A. *107*, 12941–12945. https://doi.org/10.1073/pnas.1007783107

Diamond, J. (1984). ' Normal' extinctions of isolated populations. In Extinctions, Nitecki, M.H., ed. (University of Chicago Press, Chicago), pp. 191–246.

Doolittle, W.F. (1999). Phylogenetic classification and the universal tree. Science *284*, 2124–2129.

Doolittle, W.F., and Brunet, T.D. (2016). what is the tree of life? PLOS Genet. *12*, e1005912. https://doi.org/10.1371/journal.pgen.1005912

Draper, J.L., Hansen, L.M., Bernick, D.L., Abedrabbo, S., Underwood, J.G., Kong, N., Huang, B.C., Weis, A.M., Weimer, B.C., van Vliet, A.H., *et al.* (2017). Fallacy of the unique genome: sequence diversity within single *Helicobacter pylori* strains. MBio *8*, e02321-16.

Ebrahimi-Rad, M., Bifani, P., Martin, C., Kremer, K., Samper, S., Rauzier, J., Kreiswirth, B., Blazquez, J., Jouan, M., van Soolingen, D., *et al.* (2003). Mutations in putative mutator genes of *Mycobacterium tuberculosis* strains of the W-Beijing family. Emerging Infect. Dis. *9*, 838–845.

Eigen, M., and Schuster, P. (1977). A principle of natural self-organization. Naturwissenschaften *64*, 541–565.

Eigen, M., and Schuster, P. (2012). The Hypercycle: A Principle of Natural Self-organization (Springer Science & Business Media, Berlin).

Elkins, C.A., Kotewicz, M.L., Jackson, S.A., Lacher, D.W., Abu-Ali, G.S., and Patel, I.R. (2013). Genomic paradigms for food-borne enteric pathogen analysis at the USFDA: case studies highlighting method utility, integration and resolution. Food Addit. Contam. Part A Chem. Anal. Control Expo. Risk Assess. *30*, 1422–1436. https://doi.org/10.1080/19440049.2012.743192

Green, J.L., Holmes, A.J., Westoby, M., Oliver, I., Briscoe, D., Dangerfield, M., Gillings, M., and Beattie, A.J. (2004). Spatial scaling of microbial eukaryote diversity. Nature *432*, 747–750.

Horner-Devine, M.C., Lage, M., Hughes, J.B., and Bohannan, B.J. (2004). A taxa-area relationship for bacteria. Nature *432*, 750–753.

Hu, Y., and Zhu, B. (2016). The human gut antibiotic resistome in the metagenomic era: progress and perspectives. Infect. Dis. Transl. Med. *2*, 41–47.

Hubbard, T., Barker, D., Birney, E., Cameron, G., Chen, Y., Clark, L., Cox, T., Cuff, J., Curwen, V., Down, T., *et al.* (2002). The Ensembl genome database project. Nucleic Acids Res. *30*, 38–41.

Hug, L.A., Baker, B.J., Anantharaman, K., Brown, C.T., Probst, A.J., Castelle, C.J., Butterfield, C.N., Hernsdorf, A.W., Amano, Y., Ise, K., *et al.* (2016). A new view of the tree of life. Nat. Microbiol. *1*, 16048. https://doi.org/10.1038/nmicrobiol.2016.48

Jackson, S.A., Patel, I.R., Barnaba, T., LeClerc, J.E., and Cebula, T.A. (2011). Investigating the global genomic diversity of *Escherichia coli* using a multi-genome DNA microarray platform with novel gene prediction strategies. BMC Genomics *12*, 349. https://doi.org/10.1186/1471-2164-12-349

Jacobsen, A., Hendriksen, R.S., Aaresturp, F.M., Ussery, D.W., and Friis, C. (2011). The *Salmonella enterica* pan-genome. Microb. Ecol. *62*, 487–504. https://doi.org/10.1007/s00248-011-9880-1

Jolley, K.A., Chan, M.S., and Maiden, M.C. (2004). mlstdbNet - distributed multi-locus sequence typing (MLST) databases. BMC Bioinformatics *5*, 86. https://doi.org/10.1186/1471-2105-5-86

Jolliffe, I. (2002). Principal component analysis (Wiley Online Library).

Jones, B.A., Lessler, J., Bianco, S., and Kaufman, J.H. (2015). Statistical mechanics and thermodynamics of viral evolution. PLOS ONE *10*, e0137482. https://doi.org/10.1371/journal.pone.0137482

Kaufman, J., Brodbeck, D., and Melroy, O. (1998). Critical biodiversity. Cons. Biol. *12*, 521–532.

Kittel, C., and Holcomb, D.F. (1967). Introduction to solid state physics. Am. J. Phys. *35*, 547–548.

Kumar, S., Stecher, G., and Tamura, K. (2016). MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets. Mol. Biol. Evol. *33*, 1870–1874. https://doi.org/10.1093/molbev/msw054

Leinonen, R., Sugawara, H., and Shumway, M. (2010). The sequence read archive. Nucleic Acids Res. *39*, D19–D21.

Linkenkaer-Hansen, K., Nikouline, V.V., Palva, J.M., and Ilmoniemi, R.J. (2001). Long-range temporal correlations and scaling behavior in human brain oscillations. J. Neurosci. *21*, 1370–1377.

Linz, B., Windsor, H.M., McGraw, J.J., Hansen, L.M., Gajewski, J.P., Tomsho, L.P., Hake, C.M., Solnick, J.V., Schuster, S.C., and Marshall, B.J. (2014). A mutation burst during the acute phase of *Helicobacter pylori* infection in humans and rhesus macaques. Nat. Commun. *5*, 4165. https://doi.org/10.1038/ncomms5165

Locey, K.J., and Lennon, J.T. (2016). Scaling laws predict global microbial diversity. Proc. Natl. Acad. Sci. U.S.A. *113*, 5970–5975. https://doi.org/10.1073/pnas.1521291113

Lovejoy, T.E., Rankin, J.M., Bierregaard, R., Brown, K.S., Emmons, L.H., and Van der Voort, M.E. (1984). Ecosystem decay of Amazon forest remnants. In Extinctions, Nitecki, M.H., ed. (University of Chicago Press, Chicago), pp. 295–325.

Lovejoy, T.E., Bierregaard, R.O., Rylands, A.B., Malcolm, J.R., Quintela, C.E., Harper, L.H., Brown, K.S., Powell, A.H., Powell, G.V.N., Schubart, H.O.R., *et al.* (1986). Edge and other effects of isolation on Amazon forest fragments. In Conservation Biology: The Science of Scarcity and Diversity, Soulé, M.E., ed. (Sinauer, Sunderland, MA), pp. 257–285.

Lüdeke, C.H., Kong, N., Weimer, B.C., Fischer, M., and Jones, J.L. (2015). Complete genome sequences of a clinical isolate and an environmental isolate of *Vibrio parahaemolyticus*. Genome Announc. *3*, e00216–15. https://doi.org/10.1128/genomeA.00216-15

MacArthur, R.H., and Wilson, E.O. (1963). An equilibrium theory of insular zoogeography. Evolution, 373–387.

Makarova, K., Slesarev, A., Wolf, Y., Sorokin, A., Mirkin, B., Koonin, E., Pavlov, A., Pavlova, N., Karamychev, V., Polouchine, N., *et al.* (2006). Comparative genomics of the lactic acid bacteria. Proc. Natl. Acad. Sci. U.S.A. *103*, 15611–15616.

Mandelbrot, B.B. (1983). The Fractal Geometry of Nature, Vol 173 (Macmillan, London).

Meier-Kolthoff, J.P., Auch, A.F., Klenk, H.P., and Göker, M. (2013). Genome sequence-based species delimitation with confidence intervals and improved distance functions. BMC Bioinformatics *14*, 60. https://doi.org/10.1186/1471-2105-14-60

Parkhill, J., Wren, B.W., Mungall, K., Ketley, J.M., Churcher, C., Basham, D., Chillingworth, T., Davies, R.M., Feltwell, T., Holroyd, S., *et al.* (2000). The genome sequence of the food-borne pathogen *Campylobacter jejuni* reveals hypervariable sequences. Nature *403*, 665–668. https://doi.org/10.1038/35001088

Parks, D.H., Rinke, C., Chuvochina, M., Chaumeil, P.A., Woodcroft, B.J., Evans, P.N., Hugenholtz, P., and Tyson, G.W. (2017). Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. Nat. Microbiol. *2*, 1533–1542. https://doi.org/10.1038/s41564-017-0012-7

Patel, I.R., Gangiredla, J., Lacher, D.W., Mammel, M.K., Jackson, S.A., Lampel, K.A., and Elkins, C.A. (2016). FDA *Escherichia coli* Identification (FDA-ECID) microarray: a pangenome molecular toolbox for serotyping, virulence profiling, molecular epidemiology, and phylogeny. Appl. Environ. Microbiol. *82*, 3384–3394. https://doi.org/10.1128/AEM.04077-15

Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K.S., Manichanh, C., Nielsen, T., Pons, N., Levenez, F., Yamada, T., *et al.* (2010). A human gut microbial gene catalogue established by metagenomic sequencing. Nature *464*, 59–65. https://doi.org/10.1038/nature08821

Ruggiero, M.A., Gordon, D.P., Orrell, T.M., Bailly, N., Bourgoin, T., Brusca, R.C., Cavalier-Smith, T., Guiry, M.D., and Kirk, P.M. (2015). A higher level classification of all living organisms. PLOS ONE *10*, e0119248. https://doi.org/10.1371/journal.pone.0119248

Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. Bioinformatics *30*, 2068–2069. https://doi.org/10.1093/bioinformatics/btu153

Shapiro, B.J., Friedman, J., Cordero, O.X., Preheim, S.P., Timberlake, S.C., Szabó, G., Polz, M.F., and Alm, E.J. (2012). Population genomics of early events in the ecological differentiation of bacteria. Science *336*, 48–51. https://doi.org/10.1126/science.1218198

Simpson, J.T., Wong, K., Jackman, S.D., Schein, J.E., Jones, S.J., and Birol, I. (2009). ABySS: a parallel assembler for short read sequence data. Genome Res. *19*, 1117–1123. https://doi.org/10.1101/gr.089532.108

Smalley, R., Turcotte, D.L., and Solla, S.A. (1985). A renormalization group approach to the stick-slip behavior of faults. J. Geophys. Res. Solid Earth *90*, 1894–1900.

Stern, A., Bianco, S., Yeh, M.T., Wright, C., Butcher, K., Tang, C., Nielsen, R., and Andino, R. (2014). Costs and benefits of mutational robustness in RNA viruses. Cell Rep. *8*, 1026–1036. https://doi.org/10.1016/j.celrep.2014.07.011

Suerbaum, S., and Josenhans, C. (2007). *Helicobacter pylori* evolution and phenotypic diversification in a changing host. Nat. Rev. Microbiol. *5*, 441–452.

Tang, C., and Bak, P. (1988). Critical exponents and scaling relations for self-organized critical phenomena. Phys. Rev. Lett. *60*, 2347–2350. https://doi.org/10.1103/PhysRevLett.60.2347

Tateno, Y., Imanishi, T., Miyazaki, S., Fukami-Kobayashi, K., Saitou, N., Sugawara, H., and Gojobori, T. (2002). DNA Data Bank of Japan (DDBJ) for genome scale research in life science. Nucleic Acids Res. *30*, 27–30.

Walk, S.T., Alm, E.W., Gordon, D.M., Ram, J.L., Toranzos, G.A., Tiedje, J.M., and Whittam, T.S. (2009). Cryptic lineages of the genus *Escherichia*. Appl. Environ. Microbiol. *75*, 6534–6544. https://doi.org/10.1128/AEM.01262-09

Weimer, B.C. (2012). 100K Pathogen Genome Project,. NCBI Bioproject PRJNA186441.

Weimer, B.C. (2107). 100K Pathogen Genome Project. Genome Announc. *5*, e00594–00517.

Weis, A.M., Storey, D.B., Taff, C.C., Townsend, A.K., Huang, B.C., Kong, N.T., Clothier, K.A., Spinner, A., Byrne, B.A., and Weimer, B.C. (2016). Genomic comparison of *Campylobacter* spp. and their potential for zoonotic transmission between birds, primates, and livestock. Appl. Environ. Microbiol. *82*, 7165–7175.

Weitz, J.S., Hartman, H., and Levin, S.A. (2005). Coevolutionary arms races between bacteria and bacteriophage. Proc. Natl. Acad. Sci. U.S.A. *102*, 9535–9540.

Wilson, E.O. (1999). The Diversity of Life (WW Norton & Company, New York).

Woese, C.R. (2004). A new biology for a new century. Microbiol. Mol. Biol. Rev. *68*, 173–186. https://doi.org/10.1128/MMBR.68.2.173-186.2004

Wu, D., Hugenholtz, P., Mavromatis, K., Pukall, R., Dalin, E., Ivanova, N.N., Kunin, V., Goodwin, L., Wu, M., Tindall, B.J., *et al.* (2009). A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea. Nature *462*, 1056–1060. https://doi.org/10.1038/nature08656

Xiao, Y., Rouzine, I.M., Bianco, S., Acevedo, A., Goldstein, E.F., Farkov, M., Brodsky, L., and Andino, R. (2016). RNA recombination enhances adaptability and is required for virus spread and virulence. Cell Host Microbe *19*, 493–503. https://doi.org/10.1016/j.chom.2016.03.009