Diversity within species: interpreting strains in microbiomes

Thea Van Rossum[1], Pamela Ferretti[1], Oleksandr M. Maistrenko[1] and Peer Bork[#1,2,3,4]

[1] European Molecular Biology Laboratory, Structural and Computational Biology Unit, Heidelberg, Germany
[2] Max Delbrück Centre for Molecular Medicine, Berlin, Germany
[3] Molecular Medicine Partnership Unit, University of Heidelberg and European Molecular Biology Laboratory, Heidelberg, Germany
[4] Department of Bioinformatics, Biocenter, University of Würzburg, Würzburg, Germany

[#]*e-mail*: *bork@embl.de*

Abstract

Studying within-species variation has traditionally been limited to culturable bacterial isolates and low-resolution microbial community fingerprinting. Metagenomic sequencing and technical advances have enabled culture-free, high-resolution strain and subspecies analyses at high throughput and in complex environments. This holds great scientific promise but has also led to an overwhelming number of methods and terms to describe infraspecific variation. This Review aims to clarify these advances by focusing on the diversity within bacterial and archaeal species in the context of microbiomics. We cover foundational microevolutionary concepts relevant to population genetics and summarise how within-species variation can be studied and stratified directly within microbial communities with a focus on metagenomics. Finally, we describe how common applications of within-species variation can be achieved using metagenomic data. We aim to guide the selection of appropriate terms and analytical approaches to facilitate researchers in benefiting from the increasing availability of large, high-resolution microbiome genetic sequencing data.

Introduction

For over a century, bacterial cultivation has enabled the isolation and classification of thousands of bacterial strains. Through these efforts, a species concept was translated in the bacterial context as a group of individuals that form a coherent genomic cluster[1] (see below for details and disagreements). Despite this genetic similarity, it was also established that a

large magnitude of phenotypic variance is possible among strains from the same species (’conspecific’ strains). The importance of variability within species has been particularly well studied in the context of pathogenicity, and many species have been found to have both pathogenic and commensal strains (for example, *Escherichia coli*[2] and *Bacteroides fragilis*[3]). Indeed, a classic example are *E. coli* strains, which can be pathogenic, commensal, host-associated or environmental[2]. The relationship between strain identity and host health demonstrates how it can be insufficient to study microbial communities at species level resolution, and the same applies in many other areas, such as drug response[4], nutrient cycling[5], nitrogen fixation[6] and host association[7].

Cultivation-based approaches have a fundamental[8] and continued[9] role in studying within-species variation but, despite their recent methodological progress[10], they have important limitations. Few microorganisms can be easily cultivated under isolated, laboratory conditions, and cultivation is typically low-throughput. Even when culturing is possible, organisms are then studied in isolation and not their natural community setting. Culture-free, strain-level analysis of entire microbiomes has been possible for over 15 years[11–16], but it has been limited due to shallow read depths and small sample sizes. With recent technological and algorithmic innovations in metagenomics (**Box 1**) and decreasing sequencing costs, large-scale metagenomic analysis of variation within species has become feasible. There is great promise in these approaches[17–20] and they have vastly increased the rate of discovery, but they are also leading to scientific and semantic challenges.

In the traditional cultivation approach, ’strain’ refers to a pure culture or isolate, denoting a taxonomic entity rather than a natural concept[21]. This operational definition cannot be transferred directly to the modern culture-free approaches and a widely accepted, biologically meaningful definition of strain remains elusive. Exacerbating this situation, and perhaps in response to the lack of generally accepted terminology, a plethora of overlapping terms have been coined in high-resolution microbiome studies and are often poorly defined. The resulting confusion impedes communication and synergy among researchers both in microbiome fields and beyond. To place new operational definitions in the correct context of existing conceptual definitions of within-species variation, it is essential to understand the microevolutionary processes that create and constrain variation within species.

In this Review, we summarise the processes that produce and constrain variation within species, and describe how the balance of these forces shapes the magnitude and structure of the variation. We then provide an overview of the major ways in which this variation can be studied and stratified into categories using metagenomic data and define commonly used terminology, which we then put into the context of applications. We use 'within-species variant' to refer to any grouping below the species level. Throughout this Review, we highlight the advances and challenges that are resulting from the use of metagenomic data to study within-species diversity.

Variation and cohesion within species

Processes leading to within-species variation.

Diversity within species is the result of continuous processes of variation generation and subsequent selection and drift (**Figure 1**). Mutations and gene flow introduce genetic variability into otherwise identical lineages of clonal daughter cells.

Mutations (that is, substitutions, insertions, deletions and inversions) arise continuously in the genome due to errors in the DNA replication process, damages caused by mutagens or errors in the DNA repair and recombination mechanisms[22]. Although the typical mutation rate for double-helix DNA-based organisms is approximately 1 nucleotide change per 10^9 nucleotides per replication[23], mutation rates can vary across and within species by orders of magnitude[24]. Selection for lower or higher rates balances the metabolic cost of reducing mutation frequency versus the impact of deleterious mutations[25]. The direction of this balancing depends on habitat conditions, population size and mutator allele strength[25]. The rate of accumulation of mutations within a lineage of bacteria depends on the mutation rate, as well as natural selection and genetic drift that act upon the mutations. This further diversifies the observed rates of mutation. For example, non-lethal rates of mutation have been observed from 10^{-9} to 10^{-3} mutations per genome per generation in *Vibrio* species[26,27]. Further, not all portions of the bacterial genome are equally subject to mutations. Mutation accumulation rates are higher in accessory genes than in core genes, unless a core gene is located near accessory genes or mobile genetic elements, and higher in secondary chromosomes than in primary chromosomes[28,29]. In general, deletions are more frequent than insertions and non-

functional sequences are readily lost from bacterial genomes[30,31]. Mutations that arise in one genome can be passed vertically to descendants or horizontally to neighbouring cells.

The transfer of genetic variation from one population to another (gene flow) can cause rapid and large-scale additions and rearrangements of genomic regions[32]. DNA can be transferred between cells by horizontal gene transfer (HGT) via transformation, transduction, conjugation, gene transfer agents and membrane vesicles[33,34]. Newly acquired donor DNA can stay separate within the acceptor cell (for example, as a plasmid or lytic phage) or can be incorporated into the genome of the acceptor through a number of mechanisms[34], including homologous recombination [35]. HGT is more frequent within species, but it can also occur between species[36]. It can result in replacement of genetic segments with donor homologs, often within species via homologous recombination, or in acquisition of new genetic material. In terms of impact on within-species variation, the most important factor of HGT is not the mechanism (for example, homologous recombination) but rather whether or not the genetic material being transferred is novel to the recipient population or species (discussed below). The main processes limiting HGT include lack of surface compatibility for the conjugative process, CRISPR-mediated microbial immunity[37] and restricted host specificity of bacteriophages[34]. Notorious examples of HGT between conspecific **[G]** variants include two cases where toxin genes were transferred from toxigenic to non-toxigenic strains in *Clostridioides difficile* [38] and in *E. coli*[39], with the latter causing 54 deaths in 2011 in Germany.

Natural selection and genetic drift determine the fate of within-species variation introduced through mutation and gene flow. Genetic drift randomly eliminates genetic variations within a population, whereas natural selection maintains or eliminates variations that confer a fitness advantage or disadvantage, respectively. In this context, the effect of natural selection is limited by the background noise of genetic drift[40]. Natural selection is driven by a multitude of biotic and abiotic factors that differentially influence the survival and replicative capability of species subpopulations (**Figure 1**). These factors can shape the composition of microbial communities at the species and within-species levels through community assembly[41] and classic evolutionary forces. Selective pressure factors vary from habitat to habitat and can include pH, temperature, oxygen and other gas concentration, nutrient availability, direct competition or commensalism with other bacteria, predation by phages and eukaryotes, and

presence of stress-inducing xenobiotics such as drugs, anti-microbial compounds and heavy metals.

Species definitions and mechanisms of species cohesion.

With the vertical accumulation of mutations and the horizontal acquisition of genes, variation among the descendants of one cell could constantly increase, creating a continuous landscape of genetic variation across bacterial genomes. However, when genomic similarities are compared across bacteria, distinct clusters are observed. These clusters are thought of as species in bacteria[42], though the applicability of a 'species' concept is contested[43]. In this Review, we use the word 'species' to reflect these clusters of genetic similarity.

For many decades, bacterial species delineation based on genome similarity has been measured using DNA–DNA hybridization (DDH). According to the bacterial nomenclature code, conspecific genomes have ≥70% similarity by DDH. Increasingly, DDH is complemented or replaced by DNA sequencing of isolates and average nucleotide identity (ANI) comparisons[8,44], with approximately ≥70% similarity in DDH corresponding to ≥94% of ANI in the core genome and ≥96% in universal marker genes [7,45–49]. The approximation in these correspondences can affect classification, as in the case of *Fusobacterium nucleatum*, for which subspecies were defined based on DNA–DNA hybridization[50] but then suggested to be reclassified as separate species after reassessment with *in silico* measurements of ANI[51]. As suggested by early studies[13,52,53], the presence of a distinctive bacterial species boundary is identifiable using metagenomic data and was recently confirmed by large-scale studies, which identified this boundary at ANI thresholds based on whole genomes (~95%)[54,55] and on markers genes (96.5%)[48,56] and also described a drastic drop in gene flow in core genomes.

Despite the overall consistency of genomic ANI data, defining bacterial and archaeal species remains controversial, with over twenty conceptual definitions of 'species'[57–60] and some researchers questioning the concept altogether[43]. The biological and the phylogenetic concepts of species are the most applicable for bacteria and archaea[61]. The former defines species as a group of individuals that can interbreed resulting in viable offspring, which translates to the possibility for homologous recombination in Prokaryotes, whereas the latter defines species as clades that are characterized by distinctive phenotypic properties. Both

concepts predict a decline in the homologous recombination[36,62] and HGT[63] rates between different species. The multitude of potential species definitions are not necessarily well served by ANI-based genome comparisons alone. Instead, other methods can be used to operationally define species, in addition to, or in place of ANI, such as by phenotype, similarity in universal-single copy marker genes (for example, 16S rRNA), and gene content[46,64].

The genomic similarity within species is called 'cohesion'. This is maintained predominantly through within-species recombination and selection against lower-fitness alleles[55,65]. If an allele is more beneficial than all others in a population, it can spread completely through that population, resulting in a 'hard selective sweep' [66,67]. When recombination rates are low, it is likely that the whole genome will hitchhike to prevalence along with this adaptive allele, resulting in a 'genome-wide selective sweep'[68]. When hard, whole-genome selective sweeps occur, they can reduce diversity within a species and maintain dissimilarity between species[65,69,70].

Determinants of magnitude and structure of variation within a species.

Diversity within species is generated, maintained and purged to different extents, such that some species are highly heterogenous whereas others are tightly cohesive. These features of within-species variation depend on the populations observed (**Box 3**) and can be described globally or locally. The balance between the forces that increase diversity and those that maintain cohesion shapes both the magnitude and the structure of variation within a species.

The amount of variation generated within a species depends on the mutation rate, generation time, tendency for inter-species HGT, and population size, whereas the amount of variation that persists depends on the stringency of selective pressures in its habitats, the population size[71], and the frequency and severity of selective sweeps. The balance between divergence and cohesion is modulated by selection and drift, which are shaped by biotic and abiotic factors of the ecological niche (**Figure 1**). HGT can increase the genetic variation within a population if the material being transferred is novel to the receiving population, for example if the donor cell was dispersed from a foreign population or is distantly related. Conversely, HGT can homogenise a population in terms of specific gene content or single nucleotide

variant (SNV) presence if it spreads this genetic material throughout the population, resulting in a gene-specific hard selective sweep [72].

Within a species, a structured population can arise due to a combination of soft selective sweeps — when multiple alternative adaptive alleles spread and coexist in a population[73]— along with drift and dispersal into new locations with similar or new ecological niches. For instance, when the rate of mutation generation is high and the rate of within-species recombination is low, strains may diverge into subgroups that are more internally cohesive relative to one another. Specifically, reduction of the ratio between recombination to mutation (r/m) events below 0.25 seems to enable subpopulations to diverge freely[36]. This may result in the establishment of subspecies[74,75], which are groups of strains with partially disrupted gene flow that might be in the process speciation.

Sub-speciation can be caused or accelerated by physical or geographic barriers that block gene flow between sub-speciating groups ('allopatric'), which leads to divergence of subspecies either due to natural selection or drift[76]. However, sub-speciation can also happen without spatial separation ('sympatric'). In this case, it is likely that there is a selective advantage to specialization, for example, to diminish competition for resources. Due to the extreme dispersibility of bacteria and archaea, complete physical blocks to gene flow may be rare and there might be in-between scenarios. When occasional gene flow occurs and niches overlap, purifying selection can maintain partial cohesion between subspecies, which can prevent divergence from establishing stable subspecies[77].

At one extreme, species can be monotypic; that is they have a uniform or 'smeared' distribution of genetic similarities across their entire population. Monotypic species with low diversity are more likely to be specialists, with narrow geographic distributions or host ranges, or are the product of recent speciation[78,79]. *Chlamydia trachomatis* is an example of a monotypic low diversity intracellular pathogenic species[80]. At the other extreme, species with subspecies ('polytypic') and high diversity are more likely to be free-living generalists with multiple adaptations to distinct and fluctuating environments, with broad geographic ranges or many partially overlapping niches[77,79]. For example, *E. coli* has at least six phylogroups that tend to be more prevalent in different habitats[81].

Much of the fundamental knowledge described above was obtained on a species-by-species basis through culture- and isolation-based experiments. The rise of microbiomic approaches enables the characterisation of variation across many species at a large scale and offers promising new research avenues (**Box 2**). To meaningfully place these new findings into context it is important to adapt concepts and terminology from this body of knowledge appropriately for use in metagenomic studies.

Stratification of within-species variation

Within-species variation often needs to be stratified into meaningful groups to be studied and associated with categorical variables, such as health status, geographic location or metabolic capability. The theory described above can support conceptual definitions of such groups, but these generally cannot be used directly in microbiological studies. Instead, operational definitions of variant groups must be devised based on criteria that can be measured. Typically, this is done on genetic or phenotypic scales. The appropriate metrics to use to operationally define variant groups, such as 'strains', depends on the biological questions being asked and the methodology being used (**Figure 2A**).

Genetic stratification using metagenomic data.

Within-species genetic variation can be measured in many ways, some common metrics being overall genome similarity, the number of shared and unique genes, and/or the number and nature of SNVs. In this section, we discuss how these measures are taken, and explore their strengths and limitations. When these analytical approaches are applied to the large amount of data produced by metagenomic sequencing, within-species profiling can be performed in a high-throughput manner simultaneously for many species (see, for example, Refs.[75,82–90] and examples below). However, this also raises various data quality issues, such as incomplete and partially erroneous data, as well as technical challenges, such as large computational and storage requirements.

Overall genome similarity at infra-specific levels can be assessed from metagenomic data either directly from reads and reference genomes[91–93] or through comparisons of metagenomic assembled genomes (MAGs)[54]. Reference-genome based approaches can be limited by low availability of appropriate reference genomes, especially in non-human

microbiomes. Large sets of MAGs are now available and methods to calculate ANI have improved in efficiency[94]. However, calculating ANI for large genomic cohorts remains computationally challenging[54]. Further, using MAGs in ANI comparisons can introduce inaccuracies due to data quality limitations and incompleteness (**Box 3**).

Decline of ANI and recombination rate can be indicators of ongoing subdivision of a species[57]. However, in contrast to species boundaries, within-species variants do not seem to display a universal threshold based on genome or marker genes that would categorise them into groups. Instead, the range and distribution of ANI values within species varies by taxon and population[54], which limits its utility for broad stratification. Further, genetic differences that are coded by a small number of nucleotides relative to the size of the genome, and thus have a small impact on ANI, can have a very large impact on phenotype. Therefore, at the small scale of ANI differences that occur within species, measures of gene content, SNVs and indels are more informative than ANI for defining biologically relevant within-species variants.

Gene content is the sum of all genes in a genome, including core genes (which are present in almost all conspecific variants) and accessory genes (which are only present in a subset). Differences in accessory gene content between variants can arise at the single-gene level[95] or at the genetic-segment level[82], which can include multiple genes ('structural variation'). Gene content differences can be calculated based either on the presence or absence of a gene[96], or additionally on the number of copies of that gene[97]. Gene order ('synteny') is considered within structural variation, but has not yet been addressed directly by metagenomic methods. Metagenomic data can be used to study within-species gene content variation by looking for gene clusters[95] or by associating gene content with variants defined by SNV profiles[75,98]. The relationship between gene content similarity and phylogeny is complicated by HGT. However, comparative studies of conspecific genomes have shown that pairwise similarity based on gene content is correlated with pairwise similarities based on core genome ANI[99,100] (**Figure 2B**), and that distinct SNV profiles can correspond to distinct gene profiles[75].

SNV differences can be used to compare conspecific variants at high resolution. These comparisons can consider the number of variant positions, their locations (for example, in core genes, accessory genes or intergenic regions), their spread across the genome (clustered

or disperse) and their potential phenotypic impact (for example, synonymous or non-synonymous mutations). In metagenomes, the identification of SNVs can be *de novo*[98,101], based on MAGs[102], or based on pre-existing reference genes or genomes[103–105]. The degree of similarity between the references and the actual community members can have a big impact on the accuracy of the results[106]. Identifying SNVs based on MAGs can reveal population dynamics, such as hard and soft selective sweeps in populations of lake bacteria[102], but can also introduce errors due to the potential low quality of MAG references (**Box 3**). Groups of conspecific genomes can be defined from metagenomic data based on the distinctive presence of SNVs (for example, 'SNP-types' (Ref. [107])); from thousands of SNVs indicating population structure by defining subspecies[75] and subpopulations[108], to tens of SNVs delimiting strains[107]. Isolate data has been used to show that single SNV differences can determine phenotype, such as pathogenicity[109,110] or antimicrobial drug resistance[111,112]. The ability to detect low abundance SNVs in microbiomic data is limited when sequencing depths are shallow and population sizes are large. When SNVs are likely to have been vertically transferred, then they can be used to define haplotypes and lineages. Extending this approach, SNVs can be used to reconstruct phylogeny within a species[113]; however, care must be taken to use loci that are unlikely to have been in an HGT region, such as housekeeping genes[114].

When multiple genetic variants are in one chromosome they are 'linked'. Linked variants are inherited together, but this linkage can be disrupted by recombination or mutation. Determining the linkage between alleles can be used to track lineages, reconstruct haplotypes ('phasing variants') and detect HGT. However, metagenomic data is inherently limited in providing linkage data when the typical short-read, shotgun sequencing approach is used because this method breaks up DNA. Assembling short reads may be able to recover linkage information; however chimerism is common when there are multiple highly similar genomes within one sample, such as multiple conspecific strains ('strain heterogeneity'). Instead of exact profiles of linked alleles, shotgun metagenomics is usually limited to providing sets of multiallelic loci with allele frequency information. These can still be useful for many applications, as described in the final section of this Review. They can also be used to perform population genetic analyses for a species, such as to calculate estimates of population diversity (for example, π (Pi) – diversity or average pairwise genetic difference (between individuals)), population structure (for example, fixation index (Fst) or allele

similarity between populations) and selection pressure (dN/dS, pN/pS, Tadjima's D, or Fay and Wu's H)[115].

Many software tools have been developed to measure and categorise diversity within species using metagenomic data. Generally, these have two broad aims: classification and discovery. Classification-oriented tools (for example, metaMLST)[116], PathoScope[93], MetaPhlAn2[117], StrainSifter[118], Sigma[92], SPARSE[91] StrainEst[119]) aim to detect if a known, characterised, within-species group (for example, a target genome, named strain, classic typed subspecies or MLST type) is present in a sample. Discovery-oriented tools typically group within-species variation into clusters of similarity using one of three measures: gene content (for example, PanPhlAn[95]), SNVs in whole or core genomes (for example, metaSNV[104]) or SNVs in marker genes (for example, Lineages algorithm[120], ConStrains[107], StrainPhlAn[105], DESMAN[98], StrainFinder[121], mOTUs2[56]), which might be followed up with detection of distinctive gene content (for example, DESMAN[98]). Although many tools claim to provide 'strain level' resolution, the term 'strain' is defined differently across software (see next section for discussion of definitions). The tools that can recover SNV linkage information *de novo* from SNV abundances across samples include ConStrains[107], DESMAN[98], StrainFinder[121], and the Lineages algorithm[120]. When the assumption can be made that samples contain a single dominant within-species group, tools like StrainPhlAn[105] and metaSNV[104] can also be used to cluster SNVs into within-species groups ('strains' and 'subspecies', respectively).

Although these tools enable many applications of metagenomic data to study within-species variation (see below) they have some important limitations. For example, tools that rely on mapping reads to reference genomes or marker genes are inherently limited by the availability of appropriate reference genomes, which in some environments is very low (for example, freshwater and soil). This limitation can be circumvented by building and using MAGs (for example, as in DESMAN), but MAG quality concerns must be considered, especially if time-series data is not available (**Box 3)**. Other logistical limitations include requiring an extremely high depth of coverage (for example, reported[87,98] limitation for ConStrains) and not being able to handle large magnitudes of data (for example, reported[98,104] limitation for Lineages algorithm**)**. These selected examples demonstrate how limitations of foundational software can arise as the metagenomic field progresses towards larger and more complex datasets. These and other limitations result in tools being difficult or impossible to

run, or not feasible to use with current reasonably sized datasets, preventing results from being reproducible or extendible.

The software referenced in this Review are examples of tools that reportedly perform the methodological approaches described. These references are not endorsements or reports of accuracy or usability. The reported features of many tools have been compared in recent reviews[19,20], but a thorough comparison of accuracies has not yet been completed (although are expected to be addressed in the Critical Assessment of Metagenome Interpretation (CAMI)[122] framework). Future work is expected to make comparisons for within-species analysis software; however, what exactly is meant by the specific terminology of each tool (for example, 'SNV-type', 'strain populations' etc.) and their mapping to common terms (for example, strain, subspecies) will have to be carefully handled.

Terms for genetic stratification.

There are many terms that stratify variation within species (**Table 1**). Out of the terms that are both most commonly used and recognised by the International Code of Nomenclature of Prokaryotes[44], we highlight three terms to cover the range of genetic variation within species: genome, strain and subspecies (**Figure 2C**). In this section, we discuss conflicts in the usage of these terms in culture-based microbiology and metagenomics and suggest solutions.

For decades, the most common source of microbial genomes was sequencing of isolates. Recently, this rate of production has been overtaken by rapid production of MAGs. A barrier to synergy between isolate-based and metagenomic research stems from the misinterpretation of MAGs as equivalent to isolate genomes (**Box 3**). The former might represent a population containing considerable diversity, whereas the latter usually represents a cultured isolate with little diversity. Considering also the rise in single cell sequencing, it will be useful to increasingly qualify the term genome' as: cellular, isolate, or metagenomic.

The term 'strain' is widely used across fields in microbiology and has many contrasting definitions (**Figure 2A**). In bacteriology, a strain is the "descendants of a single isolation in pure culture, and usually is made up of a succession of cultures ultimately derived from an initial single colony"[8] founded by one or more cells [44]. This is a strain in the taxonomic sense[21] ('taxonomic strain'), used for type strains and culture collections. In this case, the

origin of a strain is at isolation. An alternative definition, used for example in epidemiology, recognises a strain as an entity existing in nature[21]. This 'natural strain' is defined as a set of conspecific isolates with distinctive genotypic and/or phenotypic characteristics[123]. A 'taxonomic strain' can be thought of as an isolated, cultured sample of a 'natural strain'[21]. Operationally, the boundaries of natural and taxonomic strains vary. For example, taxonomic strains can become phenotypically heterogeneous with as few as three mutations[124], but would still be called the same strain. By contrast, in some cases, isolates need to have less than three SNV differences[125] to be considered to come from the same natural strain. This demonstrates that the genetic thresholds for strain delineation have not been universally set in culture-centric microbiology.

These two definitions of 'strain', among others[126], continue to coexist in culture-centric microbiology, and adoption of the term in microbiomics has extended this complexity. The disambiguating prefixes 'taxonomic' versus 'natural' are rarely used; however, this duality can clarify the mixed usage of the term 'strain' in metagenomics. Strain-level metagenomics often poses two types of questions: classification and discovery. Classification questions ask if genetic segments (sequencing reads) belong to a particular 'taxonomic strain', such as detecting if the probiotic strain *Bifidobacterium bifidium* BB12 is present in a stool sample. Discovery questions ask if there are subgroups within a species that form 'natural strains', for example by clustering the genetic variation of genomes or of genetic segments. Conflict can arise among metagenomic tools for strain discovery that use different definitions of a natural strain – and will implicitly therefore give different results.— for example, defining natural strains based on differential gene content[95] versus based on SNVs in shared genes[105].

A universally applicable, operational definition of strain with strong biological basis has not been established and may not exist. In theory, genomes with as few as one SNV difference could be referred to as different strains. However, this practice is not recommended due to the unmanageable number of strains it would produce from metagenomic data. There are no rules on how many SNVs define a separate strain and whether such SNVs need to be fixed in the population or need to effect phenotype. In practice, the choice of how to set this cut-off is implicit in the choice of the strain-level profiling tool (for example, more than 0.1% of the nucleotides on species-specific marker genes, as set in StrainPhlAn) or is set by the analysis authors (for example, greater than 98% ANI[127]). Given such variability in the operational

definition of strain, it becomes particularly valuable to use more specific terminology, instead of the generic term 'strain' (see Table 1 and the section entitled 'Microbiome applications of within-species variation' for guidelines).

Subspecies group conspecific strains and many definitions of the term exist[128]. In classic microbiology, subspecies are clusters of strains that are genetically or phenotypically distinct, have a type strain available[129], and are named (for example, *Bacillus subtilis* subsp. *subtilis*). Over time, the basis for classification of subspecies has shifted from qualitative phenotypic measures to genomic similarities between isolates[130]. This change has resulted in classification switches, such as the demotion of species to subspecies (for example, in *Bifidobacterium longum*[131]) and vice versa for example, in *Polynucleobacter necessarius*[132]). Thus, classic named subspecies do not (yet) necessarily align with distinct genomic clusters. By contrast, in a population biology context, a subspecies is a set of local populations that live in a subdivision of a species' spatial range and that differ from other populations of the same species[74]; for example by genotype or phenotype[128]. Adapting the term subspecies for microbiomics implies the same usage dichotomy as described for strains: *classification* of reads to an existing 'classic subspecies' and *discovery* of 'population subspecies' by clustering within-species genetic variation observed across spatial scales.

Although the strict definitions of these terms do not limit the relative amounts of variation they can each contain, in practice, it is useful to put them in context of each other and use them in the suggested ranges (**Figure 2C**). As these ranges are guidelines, actual thresholds for group delineations should be included in reports when each term is used. Importantly, 'strain' is subordinate to 'subspecies' and thus should not be used to refer generally to any grouping subordinate to species (as it sometimes is). We also discourage using the term 'sub-species' due to its different definition but visual similarity to 'subspecies'. Instead, we recommend using the terms 'infraspecific' or simply 'within-species'. For example, inappropriate usage of 'strain-level analysis' or 'sub-species analysis' would be replaced with 'infraspecific analysis' or 'within-species analysis'. Additionally, non-specific groupings within species can be referred to as 'within-species variants.

Phenotypic stratification in microbial communities.

Genetic variation within a species can manifest as phenotypic differences in complex ways. Different genetic variants can manifest as the same phenotype, whereas the same genetic variant can manifest as different phenotypes under different conditions[133]. The scale of genetic differences and their phenotypic impact are also not necessarily correlated, such as dramatically increased antibiotic resistance being conferred with as little as one SNV[111,112]. Further, different phenotypes can be observed when bacteria are cultured in isolation or in coculture or are within their natural community. For example, *Pseudomonas aeruginosa* has distinct gene expression profiles *in vitro* versus during human infection, including genes involved in antibiotic resistance, cell–cell communication and metabolism, which have implications for therapy development[134]. Differences in phenotype can also be seen within species — for example, two strains of the halophilic bacterium *Salinibacter ruber* had similar expression patterns when cultured in isolation but had distinct patterns when grown in co-culture[135]. These examples highlight the importance of studying phenotypic variation within species directly in microbiomes, and several methods exist (**Box 1**). For example, metatranscriptomics has been used to reveal functional diversity between conspecific symbionts in mussels[136] and metagenomically inferred replication rates have distinguished between infraspecific subpopulations of *Citrobacter koseri* in infants[137].

The complicated relationship between genotype and phenotype implies that phenotypic classification schemes can be at odds with genetic stratifications, and specialised vocabulary exists (**Table 1**). In medicine and epidemiology, it has been useful to categorise bacteria into (possibly polyphyletic) groups based on differential pathogenicity (pathotypes) or cell-surface antigens (serotypes). For example, the enteric *E. coli* group includes both commensal and pathogenic strains, which are divided into at least seven pathotypes[45]. In ecology, groups can also be defined based on behaviour and their functional role in a community, for example, based on the type of resources exploited and the way in which they are exploited[138,139]. Species grouped this way are called 'guild s', a concept and term which could similarly be used to describe groups of strains. This kind of grouping was designed to give an appropriate resolution for the analysis of competition within ecosystems and generalisation of findings across communities. Although phenotype is the most relevant to many biological questions, it is hard to measure at large scale (though methods are progressing[4,140]). With microbiome genetic sequencing, genotypes are much easier to measure in high throughput

but linking them to phenotype is challenging as phenotype can change drastically with habitat and small genotypic differences.

Microbiome applications of within-species variation

The many scales and dimensions of variation within species reflect the wide range of biological questions that a 'within-species investigation' can address. Isolate based approaches have been used to investigate many biological questions that involve within-species variation[141,142]. With the rise of metagenomic approaches, some of the same questions can now be investigated in high throughput and for many species in the community simultaneously (with important limitations; **Box 2**; **Box 3**). Below we describe how many of the important biological applications that were pioneered using isolate-based methods can now be investigated using a metagenomics approach. We summarise common examples of such investigations into five major themes, built around key biological questions (**Figure 3**). For each theme, we summarise methodological approaches and appropriate terminology and provide examples of relevant studies or software.

Source tracking.

Where did the cells in this sample originally come from? To determine patterns of transmission or dispersal of microbial cells, their exact source population must be identified. The probability that a cell was dispersed from or is a direct descendant of a particular source population can be calculated by comparing genetic material from the target cell or population against genetic material from its potential source population or ancestors ('source tracking', 'transmission tracking' or 'lineage tracking') (**Figure 3a**). Strategies to determine source populations from metagenomic data include detecting the presence of shared SNVs[87,88,90,143], CRISPR signals[144], or strain-specific genes[89] and genome reconstruction[145]. These approaches have been used to assess, for example, whether there is transmission of bacterial cells from the human oral cavity to the gut[87], from mother to infant[88,143], from probiotic treatment to the consumer[89], or from faecal microbiome transplant (FMT) donor to recipient[90]. These strategies can be complicated by metagenomic disruption of allele linkage, multiple source populations, and evolution of the target population since dispersal from its source. Thus, although lineage tracking approaches can be useful for pathogen source detection[145], they can also be insufficient for epidemiological outbreak analysis[146]. In the

context of source tracking, the general term ‘strain’ could be replaced by the more specific term ‘lineage’, which can be characterised by a haplotype. Determining genomic haplotypes from metagenomic data remains a challenge[147]; however long-read sequencing of single DNA molecules provides promise as error rates decline[148,149].

Phylogeny reconstruction.

What is the evolutionary history of variants within this species? In phylogeny reconstruction (**Figure 3b**), the relative ancestry of multiple lineages within a species is inferred from genetic similarity. This similarity can be based on full genomes or genetic segments (for example, marker genes). Due to HGT and homologous recombination, the phylogeny that would be reconstructed can vary based on the loci chosen and the phylogeny of genetic segments may not reflect overall genomic phylogeny[150]. Alternatively, within-species phylogenetic studies might focus on reconstruction of the history of a particular gene or plasmid within a species. Phylogeography puts these histories in the context of observed geographic distributions[151,152]. Phylogenetic analysis using isolate genomes is well established[153] and these methods can be applied to microbial communities if high quality genomes are recovered, for example using MAGs or single amplified genomes (SAGs)[86]. However, data quality issues must be considered before this application (**Box 3**). Alternatively, a typical approach is to identify conspecific, homologous genetic segments in metagenomes (for example, through alignment to reference sequences), detect SNVs in them[56,103–105] and then infer their most probable history[105]. Groups within species can be defined based on phylogeny by cutting the resultant tree at an arbitrary level of similarity, creating ‘phylotypes’. In this context, the general term ‘strain’ could be replaced by the more specific terms *clade* or *phylotype*.

Genetic population structure description.

Does this species have distinct subpopulations and/or subspecies? Describing a species’ genetic population structure can, for example, suggest its geographic history or explain heterogeneous associations with host disease states[75]. A species’ population structure can be determined by overlaying genetic data with observational data to describe the distribution of genetic similarities between variants within and across populations[154]. A uniform structure (‘smear’) occurs when there is a smooth distribution in genetic similarity across the observed species variants. This occurs when populations of ancestral and sister clades exist; that is,

there are few unobservable (extinct or undetectable) branches within a tree (**Figure 3e**). By contrast, a 'clustered' structure occurs when there is a discontinuity across genetic similarities, enabling clades to be grouped into distinct clusters. Such a non-uniform structure is created by extinct branches within a tree (**Figure 3d**). This manifests as subpopulations, which are subsets of a whole population that have distinct frequencies of genetic variations (for example, alleles or SNVs).

Metagenomics can be used to study population genetics of species within microbiomes[19] by looking for clustering of genetic similarities across potential subpopulations. Detecting subpopulations is sensitive to sampling effort, as discontinuities in genetic similarity can be due to failure to observe intermediates (**Box 3**). Assessing such genetic similarities can be based on SNV allele frequencies in whole genomes[75,104,108], SNVs in marker genes[56,105] or gene content differences[155]. When MAGs or SAGs are produced, genome-based ANI clustering can also be used[156]. MAGs can also be used to track SNV and gene content differences, such as changes in populations of lake bacteria over time[102]. In this context, 'strain' is sometimes inappropriately used to refer to a *subpopulation* or *subspecies*. Subpopulations might be *ecotypes* if they have adapted to different niches, for example, through genome-wide sweep s instead of gene-specific sweep s[72,157].

Ecological niche inference.

Have the variants within this species adapted to different conditions? Looking at within-species variants in conjunction with their habitats can provide information about their niche specificity (**Figure 3f**). When genetic data is used to make inferences about uncharacterised habitats, this is sometimes referred to as 'reverse ecology'[158]. These inquiries often aim to identify the genetic segments (for example, genes, operons, plasmids) that are key to adapting to particular environments. Acquisition of these segments might be from vertical or horizontal transmission and thus can be in contrast with the phylogenetic history of the species. For example, a gene can rapidly become ubiquitous across populations due to frequent HGT under selective conditions (gene-specific sweep) for example, in the presence of antibiotics)[72]. A common approach to investigate these questions using metagenomic data is to look at conspecific subpopulations of cells that are known to have adapted to different conditions, for example, different human host diets[84], soil versus plant-host associated[159], or

shifts in lake water habitats[102], and identify distinctive genes[82,95,98,104]. Methods used in metagenome-wide association studies (MWAS) can also be applied here, though these are not often focused on adaptive evolution of populations[160]. In this context, the general term 'strain' could be replaced by the more specific term *ecotypes*[72].

In the example shown in **Figure 3**, 'genetic population structure' investigations would focus on the allele frequency differences between European and Asian populations to decide whether these are distinct subpopulations or belong to one continuous population. Investigations on 'ecological niche inference' would focus on the gene differences in the gut-associated microbiome species associated with different diets, regardless of whether the European and Asian populations are distinct subpopulations.

Typing.

Does this species variant belong to a previously described sub-group of the species? Typing analyses assess the presence of genetic features (for example, SNVs, genes, operons or plasmids.) of specific interest in conspecific species variants (**Figure 3c**). In this context, within-species groups are not defined based on evolutionary history or habitat ranges, but simply on the presence or absence of specific genetic features. Such features may confer habitat fitness, may be transient and may only be expressed under rare or artificial conditions, such as antimicrobial resistance genes, pathogenicity genes (for example, enteropathogenic *E. coli* (EHEC)), or flagella. In this case, HGT is a major consideration; presence of a genetic feature does not necessarily reflect phylogeny. For example, *serogroups* are potentially polyphyletic groups within a species that are defined based on the presence of cell surface antigens, which allows their epidemiological classification.

Metagenomic approaches can be used to detect the genetic features that defined a type. SNVs of known[116] or novel[104] importance can be detected based on reference sequences. Detecting the presence of type-defining genes based on homology to reference sequences is well established in metagenomics[147,161], but determining with certainty that these detected genes are present in a specific strain is more difficult due to the possibility of HGT within the community. In metagenomic data, HGT can be studied directly, with[162] or without[163] assembling genomes (reviewed in Ref [164]).

Comparative analyses of within-species variants with the same phenotype can be used to discover the specific genetic features that are associated (and may be causing) the phenotype (such as in MWAS[160]). For example, conspecific cells could be grouped into a pathogenic 'variant' based only on their presence within hosts that are displaying similar symptoms, without knowing the evolutionary relationship of the cells or their typical habitats. In this context, the general term 'strain' could be replaced by the more specific term *pathotype*.

The themes described above have traditionally been investigated using isolate genomic approaches or low-resolution molecular methods (**Box 1**). As metagenomic studies increasingly create large amounts of data, dozens of new methods have been established to investigate the same questions, often with their own novel vocabulary. Considering how these new methods map back to the fundamental biological questions they are addressing and the history of research in the area will help to control the explosion of terminology. Many studies will include a combination of these themes, but considering the fundamental units separately facilitates breaking down complex questions and selecting the most appropriate methodology and terminology.

Conclusions

Despite often being the highest resolution taxonomic category considered in microbiome surveys, species can contain extreme phenotypic variability. Studying such variability used to be relatively limited in scope, with a few key isolate-based methods and a limited pool of culturable bacteria. With the development of metagenomic sequencing, the number of species that can be studied and the number of methods that can be used have increased substantially. The possibility to stratify variation within species according to many criteria, and at many scales, has also led to a growing and frequently imprecise terminology. Understanding how the variability within a species arose and identifying the central biological question being asked can help to determine the correct terminology and methodology to use. In some cases, the most appropriate term may have an operational definition, and its details and cut-off thresholds might vary across studies. To facilitate communication and collaboration, and enable future comparative meta-studies, vocabulary that does not have strict and widely-known definitions should be avoided when possible or explicitly described both in terms of the criteria and the thresholds being used. This Review aims to guide such descriptions and

support a more informed development and application of within-species investigation techniques to metagenomic data.

Bibliography

1. Moore, W. E. C. *et al.* Report of the Ad Hoc Committee on Reconciliation of Approaches to Bacterial Systematics. *Int. J. Syst. Evol. Microbiol.* **37**, 463–464 (1987).
2. Leimbach, A., Hacker, J. & Dobrindt, U. E. coli as an all-rounder: The thin line between commensalism and pathogenicity. *Curr. Top. Microbiol. Immunol.* **358**, 3–32 (2013).
3. Pierce, J. V. & Bernstein, H. D. Genomic Diversity of Enterotoxigenic Strains of Bacteroides fragilis. *PLoS One* **11**, e0158171 (2016).
4. Maier, L. *et al.* Extensive impact of non-antibiotic drugs on human gut bacteria. *Nature* **555**, 623–628 (2018).
5. Neuenschwander, S. M., Ghai, R., Pernthaler, J. & Salcher, M. M. Microdiversification in genome-streamlined ubiquitous freshwater Actinobacteria. *ISME J.* **12**, 185–198 (2018).
6. Triplett, E. Genetics of Competition for Nodulation of Legumes. *Annu. Rev. Microbiol.* **46**, 399–428 (1992).
7. Nowrouzian, F. L., Adlerberth, I. & Wold, A. E. Enhanced persistence in the colonic microbiota of Escherichia coli strains belonging to phylogenetic group B2: role of virulence factors and adherence to colonic cells. *Microbes Infect.* **8**, 834–840 (2006).
8. Whitman, W. B. & Bergey's Manual Trust. *Bergey's Manual of Systematics of Archaea and Bacteria*. *Bergey's Manual of Systematics of Archaea and Bacteria* (Wiley, 2015).
9. Zhao, S. *et al.* Adaptive Evolution within Gut Microbiomes of Healthy People. *Cell Host Microbe* **25**, 656-667.e8 (2019).
10. Lagier, J.-C. *et al.* Culturing the human microbiota and culturomics. *Nat. Rev. Microbiol.* **16**, 540–550 (2018).
11. Tyson, G. W. *et al.* Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* **428**, 37–43 (2004).
12. Allen, E. E. *et al.* Genome dynamics in a natural archaeal population. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 1883–1888 (2007).
13. Eppley, J. M., Tyson, G. W., Getz, W. M. & Banfield, J. F. Genetic exchange across a

species boundary in the archaeal genus ferroplasma. *Genetics* **177**, 407–16 (2007).

14. Eppley, J. M., Tyson, G. W., Getz, W. M. & Banfield, J. F. Strainer: Software for analysis of population variation in community genomic datasets. *BMC Bioinformatics* **8**, 398 (2007).
15. Lo, I. *et al.* Strain-resolved community proteomics reveals recombining genomes of acidophilic bacteria. *Nature* **446**, 537–541 (2007).
16. Denef, V. J. *et al.* Proteogenomic basis for ecological divergence of closely related bacteria in natural acidophilic microbial communities. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 2383–2390 (2010).
17. Segata, N. On the Road to Strain-Resolved Comparative Metagenomics. *mSystems* **3**, e00190-17 (2018).
18. Suez, J., Zmora, N., Segal, E. & Elinav, E. The pros, cons, and many unknowns of probiotics. *Nat. Med.* **25**, 716–729 (2019).
19. Denef, V. J. Peering into the Genetic Makeup of Natural Microbial Populations Using Metagenomics. in *Population Genomics: Microorganisms* 49–75 (2018).

Comprehensive review on application of metagenomic approaches for microbial population genomics.

20. Bobay, L.-M. & Raymann, K. Population Genetics of Host-Associated Microbiomes. *Curr. Mol. Biol. Reports* **5**, 128–139 (2019).
21. Dijkshoorn, L., Ursing, B. M. & Ursing, J. B. Strain, clone and species: Comments on three basic concepts of bacteriology. *J. Med. Microbiol.* **49**, 397–401 (2000).

Compares and summarises definitions of key terminology in bacteriological (isolate-based) context.

22. Brown, T. Genomes. in *Genomes. 2nd edition.* (ed. Oxford: Wiley-Liss) (2002).
23. Alberts, B., Johnson, A., Lewis, J. & et al. *Molecular Biology of the Cell. Garland Science* (Garland Science, 2002).
24. Fijalkowska, I. J., Schaaper, R. M. & Jonczyk, P. DNA replication fidelity in Escherichia coli: a multi-DNA polymerase affair. *FEMS Microbiol. Rev.* **36**, 1105–21 (2012).
25. Denamur, E. & Matic, I. Evolution of mutation rates in bacteria. *Mol. Microbiol.* **60**, 820–827 (2006).
26. Dillon, M. M., Sung, W., Sebra, R., Lynch, M. & Cooper, V. S. Genome-Wide Biases in the Rate and Molecular Spectrum of Spontaneous Mutations in Vibrio cholerae and Vibrio fischeri. *Mol. Biol. Evol.* **34**, 93–109 (2017).

27. Strauss, C., Long, H., Patterson, C. E., Te, R. & Lynch, M. Genome-Wide Mutation Rate Response to pH Change in the Coral Reef Pathogen Vibrio shilonii AK1. *MBio* **8**, e01021-17 (2017).
28. Cooper, V. S., Vohr, S. H., Wrocklage, S. C. & Hatcher, P. J. Why Genes Evolve Faster on Secondary Chromosomes in Bacteria. *PLoS Comput. Biol.* **6**, e1000732 (2010).
29. Bobay, L.-M., Traverse, C. C. & Ochman, H. Impermanence of bacterial clones. *Proc. Natl. Acad. Sci.* **112**, 8893–8900 (2015).
30. Andersson, J. O. & Andersson, S. G. E. Pseudogenes, Junk DNA, and the Dynamics of Rickettsia Genomes. *Mol. Biol. Evol.* **18**, 829–839 (2001).
31. Mira, A., Ochman, H. & Moran, N. A. Deletional bias and the evolution of bacterial genomes. *Trends Genet.* **17**, 589–96 (2001).
32. Lawrence, J. G. & Retchless, A. C. The interplay of homologous recombination and horizontal gene transfer in bacterial speciation. in *Horizontal Gene Transfer. Methods in molecular biology.* **532**, 29–53 (Humana Press, 2009).
33. Lerner, A., Matthias, T. & Aminov, R. Potential Effects of Horizontal Gene Exchange in the Human Gut. *Front. Immunol.* **8**, (2017).
34. Thomas, C. M. & Nielsen, K. M. Mechanisms of, and Barriers to, Horizontal Gene Transfer between Bacteria. *Nat. Rev. Microbiol.* **3**, 711–721 (2005).

Reviews the major concepts and mechanisms of HGT and their implications for genome flux across populations.

35. Rocha, E. P. C., Cornet, E. & Michel, B. Comparative and evolutionary analysis of the bacterial homologous recombination systems. *PLoS Genet.* **1**, 0247–0259 (2005).
36. Fraser, C., Hanage, W. P. & Spratt, B. G. Recombination and the Nature of Bacterial Speciation. *Science* **315**, 476–480 (2007).
37. Gasiunas, G., Sinkunas, T. & Siksnys, V. Molecular mechanisms of CRISPR-mediated microbial immunity. *Cell. Mol. Life Sci.* **71**, 449–465 (2014).
38. Brouwer, M. S. M. *et al.* Horizontal gene transfer converts non-toxigenic Clostridium difficile strains into toxin producers. *Nat. Commun.* **4**, 2601 (2013).
39. Kaper, J. B. & O'Brien, A. D. Overview and Historical Perspectives. in *Enterohemorrhagic Escherichia coli and Other Shiga Toxin-Producing E. coli* 3–13 (American Society of Microbiology).
40. Hallatschek, O., Hersen, P., Ramanathan, S. & Nelson, D. R. Genetic drift at expanding frontiers promotes gene segregation. *Proc. Natl. Acad. Sci.* **104**, 19926–

19930 (2007).

41. Nemergut, D. R. *et al.* Patterns and processes of microbial community assembly. *Microbiol. Mol. Biol. Rev.* **77**, 342–56 (2013).
42. Chun, J. *et al.* Proposed minimal standards for the use of genome data for the taxonomy of prokaryotes. *Int. J. Syst. Evol. Microbiol.* **68**, 461–466 (2018).
43. Doolittle, W. F. Population Genomics: How Bacterial Species Form and Why They Don't Exist. *Curr. Biol.* **22**, R451–R453 (2012).
44. International Code of Nomenclature of Prokaryotes. *Int. J. Syst. Evol. Microbiol.* **69**, S1–S111 (2019).
45. Croxen, M. A. *et al.* Recent Advances in Understanding Enteric Pathogenic Escherichia coli. *Clin. Microbiol. Rev.* **26**, 822–880 (2013).
46. Konstantinidis, K. T. & Tiedje, J. M. Genomic insights that advance the species definition for prokaryotes. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 2567–72 (2005).
47. Richter, M. & Rosselló-Móra, R. Shifting the genomic gold standard for the prokaryotic species definition. *Proc. Natl. Acad. Sci.* **106**, 19126–19131 (2009).
48. Mende, D. R., Sunagawa, S., Zeller, G. & Bork, P. Accurate and universal delineation of prokaryotic species. *Nat. Methods* **10**, 881–884 (2013).
49. Goris, J. *et al.* DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. *Int. J. Syst. Evol. Microbiol.* **57**, 81–91 (2007).
50. Dzink, J. L., Sheenan, M. T. & Socransky, S. S. Proposal of Three Subspecies of Fusobacterium nucleaturn Knorr 1922: Fusobacterium nucleatum subsp. nucleatum subsp. nov. , comb. nov. ; Fusobacterium nucleatum subsp. polymorphum subsp. nov. , norn. rev. , comb. nov.; and Fusobacterium nucleatum subsp. vincentii subsp. nov., norn. rev., comb. nov. *Int. J. Syst. Bacteriol.* **40**, 74–78 (1990).
51. Kook, J. K. *et al.* Genome-Based Reclassification of Fusobacterium nucleatum Subspecies at the Species Level. *Curr. Microbiol.* **74**, 1137–1147 (2017).
52. Konstantinidis, K. T. & Delong, E. F. Genomic patterns of recombination clonal divergence and environment in marine microbial populations. *ISME J.* **2**, 1052–1065 (2008).
53. Caro-Quintero, A. & Konstantinidis, K. T. Bacterial species may exist, metagenomics reveal. *Environmental Microbiology* **14**, 347–355 (2012).
54. Jain, C., Rodriguez-R, L. M., Phillippy, A. M., Konstantinidis, K. T. & Aluru, S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries.

Nat. Commun. **9**, 5114 (2018).

55. Olm, M. R. *et al.* Consistent Metagenome-Derived Metrics Verify and Delineate Bacterial Species Boundaries. *mSystems* **5**, 647511 (2020).
56. Milanese, A. *et al.* Microbial abundance, activity and population genomic profiling with mOTUs2. *Nat. Commun.* **10**, 1014 (2019).
57. Mayden, R. L. A hierarchy of species concepts: the denouement in the saga of the species problem. *Species. The units of biodiversity.* 381–423 (1997).
58. Wilkins, J. S. How to be a chaste species pluralist-realist: the origins of species modes and the synapomorphic species concept. *Biol. Philos.* **18**, 621–638 (2003).
59. Hey, J. The mind of the species problem. *Trends Ecol. Evol.* **16**, 326–329 (2001).
60. Bapteste, E. *et al.* Prokaryotic evolution and the tree of life are two different things. *Biol. Direct* **4**, 34 (2009).
61. Konstantinidis, K. T., Ramette, A. & Tiedje, J. M. The bacterial species definition in the genomic era. *Philos. Trans. R. Soc. B Biol. Sci.* **361**, 1929–1940 (2006).
62. Bobay, L.-M. & Ochman, H. Biological Species Are Universal across Life's Domains. *Genome Biol. Evol.* **9**, 491–501 (2017).
63. Moldovan, M. A. & Gelfand, M. S. Pangenomic Definition of Prokaryotic Species and the Phylogenetic Structure of Prochlorococcus spp. *Front. Microbiol.* **9**, 428 (2018).
64. Snel, B., Bork, P. & Huynen, M. A. Genome phylogeny based on gene content. *Nat. Genet.* **21**, 108–110 (1999).
65. Achtman, M. & Wagner, M. Microbial diversity and the genetic nature of microbial species. *Nat. Rev. Microbiol.* **6**, 431–440 (2008).
66. Barton, N. H. The effect of hitch-hiking on neutral genealogies. *Genet. Res.* **72**, 123–133 (1998).
67. Hermisson, J. & Pennings, P. S. Soft sweeps and beyond: understanding the patterns and probabilities of selection footprints under rapid adaptation. *Methods Ecol. Evol.* **8**, 700–716 (2017).
68. Shapiro, B. J. *et al.* Population genomics of early events in the ecological differentiation of bacteria. *Science* **336**, 48–51 (2012).

Demonstrates that gene-specific selective sweeps followed by gradually decreasing gene flow can lead to ecologically differentiated conspecific subpopulations.

69. Cohan, F. M. Bacterial Species and Speciation. *Syst. Biol.* **50**, 513–524 (2001).
70. Cohan, F. M. Periodic Selection and Ecological Diversity in Bacteria. in *Selective Sweep* 78–93 (Springer US, 2007).

71. Charlesworth, B. Effective population size and patterns of molecular evolution and variation. *Nat. Rev. Genet.* **10**, 195–205 (2009).
72. Cohan, F. M. M. Bacterial Speciation: Genetic Sweeps in Bacterial Species. *Curr. Biol.* **26**, R112–R115 (2016).
73. Hermisson, J. & Pennings, P. S. Soft sweeps: molecular population genetics of adaptation from standing genetic variation. *Genetics* **169**, 2335–52 (2005).
74. Monroe, B. A Modern Concept of the Subspecies. *Auk* **99**, 608–609 (1982).
75. Costea, P. *et al.* Subspecies in the global human gut microbiome. *Mol Syst Biol* **13**, 960–960 (2017).
76. Retchless, A. C. & Lawrence, J. G. Temporal Fragmentation of Speciation in Bacteria. *Science* **317**, 1093–1096 (2007).
77. Shapiro, B. J. What Microbial Population Genomics Has Taught Us About Speciation. in *Population Genomics: Microorganisms* (eds. Polz, M. F. & Rajora Editors, O. P.) 31–47 (Springer Nature, 2018).
78. Sheppard, S. K., Guttman, D. S. & Fitzgerald, J. R. Population genomics of bacterial host adaptation. *Nat. Rev. Genet.* **19**, 549–565 (2018).

An extensive review about origins of genetic population structure in Prokaryotes and how to study it in context of host-microbiome interactions and adaptations.

79. Bobay, L.-M. & Ochman, H. Factors driving effective population size and pan-genome evolution in bacteria. *BMC Evol. Biol.* **18**, 153 (2018).
80. Smelov, V. *et al.* Chlamydia trachomatis Strain Types Have Diversified Regionally and Globally with Evidence for Recombination across Geographic Divides. *Front. Microbiol.* **8**, 2195 (2017).
81. Tenaillon, O., Skurnik, D., Picard, B. & Denamur, E. The population genetics of commensal Escherichia coli. *Nat. Rev. Microbiol.* **8**, 207–217 (2010).
82. Zeevi, D. *et al.* Structural variation in the gut microbiome associates with host health. *Nature* **568**, 43–48 (2019).
83. Lloyd-Price, J. *et al.* Strains, functions and dynamics in the expanded Human Microbiome Project. *Nature* **550**, 61–66 (2017).
84. De Filippis, F. *et al.* Distinct Genetic and Functional Traits of Human Intestinal Prevotella copri Strains Are Associated with Different Habitual Diets. *Cell Host Microbe* **25**, 444-453.e3 (2019).
85. Ferretti, P. *et al.* Mother-to-Infant Microbial Transmission from Different Body Sites Shapes the Developing Infant Gut Microbiome. *Cell Host Microbe* **24**, 133-145.e5

(2018).

86. Stewart, R. D. *et al.* Compendium of 4,941 rumen metagenome-assembled genomes for rumen microbiome biology and enzyme discovery. *Nat. Biotechnol.* **37**, 953–961 (2019).
87. Schmidt, T. S. *et al.* Extensive transmission of microbes along the gastrointestinal tract. *Elife* **8**, (2019).
88. Asnicar, F. *et al.* Studying Vertical Microbiome Transmission from Mothers to Infants by Strain-Level Metagenomic Profiling. *mSystems* **2**, (2017).
89. Zmora, N. *et al.* Personalized Gut Mucosal Colonization Resistance to Empiric Probiotics Is Associated with Unique Host and Microbiome Features. *Cell* **174**, 1388-1405.e21 (2018).
90. Smillie, C. S. *et al.* Strain Tracking Reveals the Determinants of Bacterial Engraftment in the Human Gut Following Fecal Microbiota Transplantation Article Strain Tracking Reveals the Determinants of Bacterial Engraftment in the Human Gut Following Fecal Microbiota Transplanta. *Cell Host Microbe* **23**, 229-240.e5 (2018).
91. Zhou, Z., Luhmann, N., Alikhan, N. F., Quince, C. & Achtman, M. Accurate Reconstruction of Microbial Strains from Metagenomic Sequencing Using Representative Reference Genomes. in *Research in Computational Molecular Biology. RECOMB 2018. Lecture Notes in Computer Science.* **10812 LNBI**, 225–240 (2018).
92. Ahn, T.-H., Chai, J. & Pan, C. Sigma: Strain-level inference of genomes from metagenomic analysis for biosurveillance. *Bioinformatics* **31**, 170–177 (2015).
93. Hong, C. *et al.* PathoScope 2.0: a complete computational framework for strain identification in environmental or clinical sequencing samples. *Microbiome* **2**, 33 (2014).
94. Ondov, B. D. *et al.* Mash: Fast genome and metagenome distance estimation using MinHash. *Genome Biol.* **17**, 029827 (2016).
95. Scholz, M. *et al.* Strain-level microbial epidemiology and population genomics from shotgun metagenomics. *Nat. Methods* **13**, 435–438 (2016).
96. Zhu, A., Sunagawa, S., Mende, D. R. & Bork, P. Inter-individual differences in the gene content of human gut bacterial species. *Genome Biol.* **16**, 82 (2015).
97. Greenblum, S., Carr, R. & Borenstein, E. Extensive Strain-Level Copy-Number Variation across Human Gut Microbiome Species. *Cell* **160**, 583–594 (2015).
98. Quince, C. *et al.* DESMAN: a new tool for de novo extraction of strains from

metagenomes. *Genome Biol.* **18**, 181 (2017).

99. Maistrenko, O. M. *et al.* Disentangling the impact of environmental and phylogenetic constraints on prokaryotic within-species diversity. *ISME J.* **1**, 735696 (2020).
100. Andreani, N. A., Hesse, E. & Vos, M. Prokaryote genome fluidity is dependent on effective population size. *ISME J.* **11**, 1719–1721 (2017).
101. Nayfach, S., Shi, Z. J., Seshadri, R., Pollard, K. S. & Kyrpides, N. C. New insights from uncultivated genomes of the global human gut microbiome. *Nature* **568**, 505–510 (2019).
102. Bendall, M. L. *et al.* Genome-wide selective sweeps and gene-specific sweeps in natural bacterial populations. *ISME J.* **10**, 1589–1601 (2016).
103. Nayfach, S., Rodriguez-Mueller, B., Garud, N. & Pollard, K. S. An integrated metagenomics pipeline for strain profiling reveals novel patterns of bacterial transmission and biogeography. *Genome Res.* **26**, 1612–1625 (2016).
104. Costea, P. I. *et al.* metaSNV: A tool for metagenomic strain level analysis. *PLoS One* **12**, e0182392 (2017).
105. Truong, D. T., Tett, A., Pasolli, E., Huttenhower, C. & Segata, N. Microbial strain-level population structure and genetic diversity from metagenomes. *Genome Res.* **27**, 626–638 (2017).
106. Bush, S. J. *et al.* Genomic diversity affects the accuracy of bacterial single-nucleotide polymorphism–calling pipelines. *Gigascience* **9**, giaa007 (2020).
107. Luo, C. *et al.* ConStrains identifies microbial strains in metagenomic datasets. *Nat. Biotechnol.* **33**, 1045–1052 (2015).
108. Delmont, T. O. *et al.* Single-amino acid variants reveal evolutionary processes that shape the biogeography of a global SAR11 subclade. *Elife* **8**, (2019).
109. Jackson, R. W. *et al.* Identification of a pathogenicity island, which contains genes for virulence and avirulence, on a large native plasmid in the bean pathogen Pseudomonas syringae pathovar phaseolicola. *Proc. Natl. Acad. Sci.* **96**, 10875–10880 (1999).
110. Scholz, B. K., Jakobek, J. L. & Lindgren, P. B. Restriction fragment length polymorphism evidence for genetic homology within a pathovar of Pseudomonas syringae. *Appl. Environ. Microbiol.* **60**, 1093–1100 (1994).
111. Pan, X. S., Yague, G. & Fisher, L. M. Quinolone resistance mutations in Streptococcus pneumoniae gyrA and parC proteins: Mechanistic insights into quinolone action from enzymatic analysis, intracellular levels, and phenotypes of wild-type and mutant

proteins. *Antimicrob. Agents Chemother.* **45**, 3140–3147 (2001).

112. Forslund, K., Sunagawa, S., Coelho, L. P. & Bork, P. Metagenomic insights into the human gut resistome and the forces that shape it. *BioEssays* **36**, 316–329 (2014).
113. Petkau, A. *et al.* SNVPhyl: a single nucleotide variant phylogenomics pipeline for microbial genomic epidemiology. *Microb. Genomics* **3**, e000116 (2017).
114. Jain, R., Rivera, M. C., Lake, J. A. & Lake, J. Horizontal gene transfer among genomes: The complexity hypothesis. *Proc. Natl. Acad. Sci.* **96**, 3801–3806 (1999).
115. Polz, M. F. & Rajora, O. P. *Population genomics : microorganisms*. (2019).
116. Zolfo, M., Tett, A., Jousson, O., Donati, C. & Segata, N. MetaMLST: Multi-locus strain-level bacterial typing from metagenomic samples. *Nucleic Acids Res.* **45**, e7 (2017).
117. Truong, D. T. *et al.* MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nature Methods* **12**, 902–903 (2015).
118. Tamburini, F. B. *et al.* Precision identification of diverse bloodstream pathogens in the gut microbiome. *Nat. Med.* **24**, 1809–1814 (2018).
119. Albanese, D. & Donati, C. Strain profiling and epidemiology of bacterial species from metagenomic sequencing. *Nat. Commun.* **8**, 2260 (2017).
120. O'Brien, J. D. *et al.* A Bayesian approach to inferring the phylogenetic structure of communities from metagenomic data. *Genetics* **197**, 925–37 (2014).
121. Smillie, C. S. *et al.* Strain Tracking Reveals the Determinants of Bacterial Engraftment in the Human Gut Following Fecal Microbiota Transplantation. *Cell Host Microbe* **23**, 229-240.e5 (2018).
122. Sczyrba, A. *et al.* Critical Assessment of Metagenome Interpretation - A benchmark of metagenomics software. *Nat. Methods* **14**, 1063–1071 (2017).
123. Struelens, M. J. *et al.* Consensus guidelines for appropriate use and evaluation of microbial epidemiologic typing systems. *Clin. Microbiol. Infect.* **2**, 2–11 (1996).
124. Spira, B., De Almeida Toledo, R., Maharjan, R. P. & Ferenci, T. The uncertain consequences of transferring bacterial strains between laboratories - RpoS instability as an example. *BMC Microbiol.* **11**, (2011).
125. Kong, L. Y. *et al.* Clostridium difficile: Investigating Transmission Patterns between Infected and Colonized Patients Using Whole Genome Sequencing. *Clin. Infect. Dis.* **68**, 204–209 (2019).
126. Saak, C. C. & Gibbs, K. A. The Self-Identity Protein IdsD Is Communicated between

Cells in Swarming Proteus mirabilis Colonies. *J. Bacteriol.* **198**, 3278–3286 (2016).

127. Brooks, B. *et al.* Strain-resolved analysis of hospital rooms and infants reveals overlap between the human and room microbiome. *Nat. Commun.* **8**, 1814 (2017).
128. Patten, M. A. Subspecies and the philosophy of science. *Auk* **132**, 481–485 (2015).
129. International Committee on Systematics of Prokaryotes. International Code of Nomenclature of Prokaryotes: Prokaryotic Code (2008 Revision). *Int. J. Syst. Evol. Microbiol.* **69**, S1–S111 (2019).
130. Meier-Kolthoff, J. P. *et al.* Complete genome sequence of DSM 30083(T), the type strain (U5/41(T)) of Escherichia coli, and a proposal for delineating subspecies in microbial taxonomy. *Stand. Genomic Sci.* **9**, 2 (2014).
131. Fukuyama, M. *et al.* Unification of Bifidobacterium infantis and Bifidobacterium suis as Bifidobacterium longum. *Int. J. Syst. Evol. Microbiol.* **52**, 1945–1951 (2002).
132. Hahn, M. W., Schmidt, J., Pitt, A., Taipale, S. J. & Lang, E. Reclassification of four Polynucleobacter necessarius strains as representatives of Polynucleobacter asymbioticus comb. nov., Polynucleobacter duraquae sp. nov., Polynucleobacter yangtzensis sp. nov. and Polynucleobacter sinensis sp. nov., and emended description of Polynucleobacter necessarius. *Int. J. Syst. Evol. Microbiol.* **66**, 2883–2892 (2016).
133. Ackermann, M. A functional perspective on phenotypic heterogeneity in microorganisms. *Nat. Rev. Microbiol.* **13**, 497–508 (2015).
134. Cornforth, D. M. *et al.* Pseudomonas aeruginosa transcriptome during human infection. *Proc. Natl. Acad. Sci. U. S. A.* **115**, E5125–E5134 (2018).
135. González-Torres, P. *et al.* Interactions between closely related bacterial strains are revealed by deep transcriptome sequencing. *Appl. Environ. Microbiol.* **81**, 8445–56 (2015).
136. Ansorge, R. *et al.* Functional diversity enables multiple symbiont strains to coexist in deep-sea mussels. *Nat. Microbiol.* **4**, 2487–2497 (2019).
137. Olm, M. R. *et al.* Identical bacterial populations colonize premature infant gut, skin, and oral microbiomes and exhibit different in situ growth rates. *Genome Res.* **27**, 601–612 (2017).
138. Pedrós-Alió, C. Toward an Autecology of Bacterioplankton. in *Plankton Ecology* 297–336 (Springer, Berlin, Heidelberg, 1989).
139. Root, R. B. The Niche Exploitation Pattern of the Blue-Gray Gnatcatcher. *Ecol. Monogr.* **37**, 317–350 (1967).

140. Mateus, A. *et al.* Thermal proteome profiling in bacteria: probing protein state in vivo. *Mol. Syst. Biol.* **14**, e8242 (2018).
141. Land, M. *et al.* Insights from 20 years of bacterial genome sequencing. *Funct Integr Genomics* **15**, 141–161 (2015).
142. Gutleben, J. *et al.* The multi-omics promise in context: from sequence to microbial isolate. *Critical Reviews in Microbiology* **44**, 212–229 (2018).
143. Ferretti, P. *et al.* Mother-to-Infant Microbial Transmission from Different Body Sites Shapes the Developing Infant Gut Microbiome. *Cell Host Microbe* **24**, 133-145.e5 (2018).
144. Lam, T. J. & Ye, Y. CRISPRs for Strain Tracking and Their Application to Microbiota Transplantation Data Analysis. *Cris. J.* **2**, 41–50 (2019).
145. Mu, A. *et al.* Reconstruction of the Genomes of Drug-Resistant Pathogens for Outbreak Investigation through Metagenomic Sequencing. *mSphere* **4**, (2019).
146. Didelot, X., Walker, A. S., Peto, T. E., Crook, D. W. & Wilson, D. J. Within-host evolution of bacterial pathogens. *Nat. Rev. Microbiol.* **14**, 150–162 (2016).
147. Quince, C., Walker, A. W., Simpson, J. T., Loman, N. J. & Segata, N. Shotgun metagenomics, from sampling to analysis. *Nat. Biotechnol.* **35**, 833–844 (2017).

Reviews how microbial communities can be studied using metagenomic sequencing, with comments on sources of bias and comparisons of analytical methods.

148. Koren, S. & Phillippy, A. M. One chromosome, one contig: complete microbial genomes from long-read sequencing and assembly. *Curr. Opin. Microbiol.* **23**, 110–120 (2015).
149. Somerville, V. *et al.* Long-read based de novo assembly of low-complexity metagenome samples results in finished genomes and reveals insights into strain diversity and an active phage system. *BMC Microbiol.* **19**, 143 (2019).
150. Jiang, X. *et al.* Dissemination of antibiotic resistance genes from antibiotic producers to pathogens. *Nat. Commun.* **8**, 15784 (2017).
151. Linz, B. *et al.* An African origin for the intimate association between humans and Helicobacter pylori. *Nature* **445**, 915–918 (2007).
152. Thorell, K. *et al.* Rapid evolution of distinct Helicobacter pylori subpopulations in the Americas. *PLoS Genet.* **13**, (2017).
153. Gardy, J. L. *et al.* Whole-Genome Sequencing and Social-Network Analysis of a Tuberculosis Outbreak. *N. Engl. J. Med.* **364**, 730–739 (2011).
154. Gregory, A. C. *et al.* Marine DNA Viral Macro- and Microdiversity from Pole to Pole.

Cell **177**, 1109-1123.e14 (2019).

155. Arevalo, P., VanInsberghe, D., Elsherbini, J., Gore, J. & Polz, M. F. A Reverse Ecology Approach Based on a Biological Definition of Microbial Populations. *Cell* **178**, 820-834.e14 (2019).
156. Garcia, S. L. *et al.* Contrasting patterns of genome-level diversity across distinct co-occurring bacterial populations. *ISME J.* **12**, 742–755 (2018).
157. Kopac, S. *et al.* Genomic Heterogeneity and Ecological Speciation within One Subspecies of Bacillus subtilis. *Appl. Environ. Microbiol.* **80**, 4842–4853 (2014).
158. Levy, R. & Borenstein, E. Reverse Ecology: From Systems to Environments and Back. in *Evolutionary Systems Biology* **751**, 329–345 (Springer, New York, NY, 2012).
159. Burghardt, L. T. *et al.* Select and resequence reveals relative fitness of bacteria in symbiotic and free-living environments. *Proc. Natl. Acad. Sci.* **115**, 2425–2430 (2018).
160. Wang, J. & Jia, H. Metagenome-wide association studies: fine-mining the microbiome. *Nat. Rev. Microbiol.* **14**, 508–522 (2016).
161. Knight, R. *et al.* Best practices for analysing microbiomes. *Nat. Rev. Microbiol.* **16**, 410–422 (2018).
162. Song, W., Wemheuer, B., Zhang, S., Steensen, K. & Thomas, T. MetaCHIP: community-level horizontal gene transfer identification through the combination of best-match and phylogenetic approaches. *Microbiome* **7**, 36 (2019).
163. Seiler, E., Trappe, K. & Renard, B. Y. Where did you come from, where did you go: Refining metagenomic analysis tools for horizontal gene transfer characterisation. *PLOS Comput. Biol.* **15**, e1007208 (2019).
164. Douglas, G. M. & Langille, M. G. I. Current and promising approaches to identify horizontal gene transfer events in metagenomes. *Genome Biol. Evol.* **11**, 2750–2766 (2019).
165. C. Barry Cox, Peter D. Moore, R. L. *Biogeography: An Ecological and Evolutionary Approach*. (2016).
166. Arora, D., Singh, A., Sharma, V., Bhaduria, H. S. & Patel, R. B. HgsDb: Haplogroups Database to understand migration and molecular risk assessment. *Bioinformation* **11**, 272–275 (2015).
167. Cantino, P. & de Queiroz, K. PhyloCode: A Phylogenetic Code of Biological Nomenclature. *PhyloCode*. www.ohiou.edu/phylocode (2010).
168. Tenover, F. C. *et al.* Interpreting chromosomal DNA restriction patterns produced by

pulsed- field gel electrophoresis: Criteria for bacterial strain typing. *J. Clin. Microbiol.* **33**, 2233–2239 (1995).

169. Schloter, M., Lebuhn, M., Heulin, T. & Hartmann, A. Ecology and evolution of bacterial microdiversity. *FEMS Microbiology Reviews* **24**, 647–660 (2000).
170. Hamilton, M. *Population Genetics*. (Wiley-Blackwell, 2009).
171. Cohan, F. M. Transmission in the Origins of Bacterial Diversity, From Ecotypes to Phyla. *Microbiol. Spectr.* **5**, (2017).
172. Alkan, C., Coe, B. P. & Eichler, E. E. Genome structural variation discovery and genotyping. *Nat. Rev. Genet.* **12**, 363–376 (2011).
173. Kaper, J. B., Nataro, J. P. & Mobley, H. L. T. Pathogenic Escherichia coli. *Nat. Rev. Microbiol.* **2**, 123–140 (2004).
174. Samuel, B. *Medical Microbiology*. (Univ of Texas Medical Branch, 1996).
175. Kenneth, R., George, R. & Sherris, J. C. *Medical microbiology : an introduction to infectious diseases*. (McGraw-Hill Medical, 2004).
176. *The American Heritage Medical Dictionary - Serovar*. (Houghton Mifflin, 2007).
177. Silva, N. A. *et al.* Genomic Diversity between Strains of the Same Serotype and Multilocus Sequence Type among Pneumococcal Clinical Isolates. *Infect. Immun.* **74**, 3513–3518 (2006).
178. Fratamico, P. M. *et al.* Advances in Molecular Serotyping and Subtyping of Escherichia coli†. *Front. Microbiol.* **7**, (2016).
179. Miller-Keane & Marie, O. Miller-Keane Encyclopedia and Dictionary of Medicine, Nursing, and Allied Health, Seventh Edition. *Saunders, an imprint of Elsevier, Inc.* (2003).
180. diCenzo, G. C. & Finan, T. M. The Divided Bacterial Genome: Structure, Function, and Evolution. *Microbiol. Mol. Biol. Rev.* **81**, (2017).
181. Hamady, M. & Knight, R. Microbial community profiling for human microbiome projects: Tools, techniques, and challenges. *Genome Res.* **19**, 1141–1152 (2009).
182. Nocker, A., Burr, M. & Camper, A. K. Genotypic Microbial Community Profiling: A Critical Technical Review. *Microb. Ecol.* **54**, 276–289 (2007).

Reviews foundational methods that enabled microbial diversity to be assessed directly within a microbial community, sometimes at within-species resolution.

183. Eren, A. M., Borisy, G. G., Huse, S. M. & Mark Welch, J. L. Oligotyping analysis of the human oral microbiome. *Proc. Natl. Acad. Sci.* **111**, E2875–E2884 (2014).
184. Eren, A. M. *et al.* Minimum entropy decomposition: Unsupervised oligotyping for

sensitive partitioning of high-throughput marker gene sequences. *ISME J.* **9**, 968–979 (2015).

185. Callahan, B. J. *et al.* DADA2: High-resolution sample inference from Illumina amplicon data. *Nat. Methods* **13**, 581–583 (2016).
186. Amir, A. *et al.* Deblur Rapidly Resolves Single-Nucleotide Community Sequence Patterns. *mSystems* **2**, (2017).
187. Tikhonov, M., Leach, R. W. & Wingreen, N. S. Interpreting 16S metagenomic data without clustering to achieve sub-OTU resolution. *ISME J.* **9**, 68–80 (2015).
188. Johnson, J. S. *et al.* Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis. *Nat. Commun.* **10**, 5029 (2019).
189. Nielsen, H. B. *et al.* Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. *Nat. Biotechnol.* **32**, 822–828 (2014).
190. Yu, F. B. *et al.* Microfluidic-based mini-metagenomics enables discovery of novel microbial lineages from complex environmental samples. *Elife* **6**, (2017).
191. Shi, X. *et al.* Microfluidics-Based Enrichment and Whole-Genome Amplification Enable Strain-Level Resolution for Airway Metagenomics. *mSystems* **4**, (2019).
192. Bowers, R. M. *et al.* Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat. Biotechnol.* **35**, 725–731 (2017).

Establishes minimal quality reporting requirements for metagenome-assembled genomes (MAGs).

193. Almeida, A. *et al.* A new genomic blueprint of the human gut microbiota. *Nature* **568**, 499–504 (2019).
194. Beitel, C. W. *et al.* Strain- and plasmid-level deconvolution of a synthetic metagenome by sequencing proximity ligation products. *PeerJ* **2**, e415 (2014).
195. Costea, P. I. *et al.* Towards standards for human fecal sample processing in metagenomic studies. *Nat. Biotechnol.* **35**, 1069–1076 (2017).
196. Shaiber, A. & Eren, A. M. Composite Metagenome-Assembled Genomes Reduce the Quality of Public Genome Repositories. *MBio* **10**, (2019).

Provides an example of how assembling genomes from metagenomes (creating MAGs) can lead to poor quality genomic data and why these genomes should not be considered the same as genomes from isolates.

197. Schmidt, T. S. B., Raes, J. & Bork, P. The Human Gut Microbiome: From Association

to Modulation. *Cell* **172**, 1198–1215 (2018).

Reviews the known connections between human gut microbiome and health, including discussion of strain-level variation.

198. Salter, S. J. *et al.* Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biol.* **12**, 87 (2014).
199. Goldstein, S., Beka, L., Graf, J. & Klassen, J. L. Evaluation of strategies for the assembly of diverse bacterial genomes using MinION long-read sequencing. *BMC Genomics* **20**, 23 (2019).
200. Alneberg, J. *et al.* Genomes from uncultivated prokaryotes: a comparison of metagenome-assembled and single-amplified genomes. *Microbiome* **6**, 173 (2018).
201. Gawad, C., Koh, W. & Quake, S. R. Single-cell genome sequencing: current state of the science. *Nat. Rev. Genet.* **17**, 175–88 (2016).

Acknowledgements

Funding for research in the authors' laboratories was provided by the European Research Council (ERC) (grant ERC-AdG-669830 MicrobioS), the European Union's Horizon 2020 Research and Innovation Programme (grant 825694 MICROB-PREDICT), the German Federal Ministry of Education and Research (Bundesministerium für Bildung und Forschung, BMBF) (grant 01GL1746B PRIMAL), and the European Molecular Biology Laboratory (EMBL).

Competing interests

The authors declare no competing interests.

Author contributions

The authors contributed equally to all aspects of the article.

Table 1. Definitions of terms used to stratify or describe variation within species.

Term	Definition	Notes
Genotype	The set of alleles of an organism.	Variable throughout time due to mutation and recombination.
Haplotype	Set of alleles or single nucleotide variants (SNVs) that are inherited together from a single parent[165].	Genetic signature of a lineage or clonal line, which can be disrupted through recombination.
Haplogroup	Group of similar haplotypes with a common ancestor that has a clade-specific SNV or SNVs[166].	In human context, used to describe a group of people that share a common ancestor.
Lineage and sublineage	Unbranched sequence of ancestral and descendant entities. Each ancestor may have multiple descendants, but only one is included in the lineage. Each entity could be an organism, clade, population or subspecies, among others. [167] A sublineage is a subsection of a lineage.	Can be visualised as an unbranched path through an evolutionary tree.
Clone	Population of bacterial cells derived from a single parent cell[129].	In evolutionary terms, it is assumed to include all the descendants of the parent cell (monophyletic)[21]. Cultured isolates are samples of clones.
Isolate	A pure culture obtained from a single colony separated from others *in vitro*[168].	Presumed to be and usually is derived from a single organism.
Clade	Group of taxonomic entities composed of one ancestor and all of its evolutionary descendants[167].	Synonym: monophyletic group.
Strain	Set of genetically similar descendants of a single colony or cell[44]. Depending on the field, it can be genetic- or phenotypic-based.	Descriptive subdivision of a species. Used widely but often with loose and/or inconsistent definitions. Can be described as 'taxonomic' or 'natural'[21].
Within-species variant	Any sub-classification of a species.	General term that does not imply a level of resolution or phylogeny.
Classic or typed subspecies	Set of strains that are genetically or phenotypically distinct and have a type strain available in a culture collection[169]; for example, *Lactococcus lactis* subspecies lactis and *L. lactis* ssp. cremoris.	The name of a classic subspecies cannot be validly published if the description is based on studies of a mixed culture[44]. Variety was used as synonym of subspecies (now deprecated)[44].
Population subspecies	Set of local populations of strains that live in a subdivision of a species' spatial range and	Species with subspecies are 'polytypic', without are

	differs from other populations of the same species by phenotypic or genotypic characteristics[74,128].	'monotypic'.
Population	Group of organisms, which live in a particular location or ecological niche at a given time.	Can be used to refer to all members of a species or to a subset of the entire population.
Subpopulation	Portion of a population that is partially isolated from others and in which allele frequencies evolve independently[170].	A 'metapopulation' is a group of subpopulations.
Strain population	A set of strains living simultaneously in the same spatial location or niche.	Distinct from population subspecies, which can include multiple populations or ecotypes.
Ecotype	An ecologically homogeneous population[72]. A clade within a species that has adapted to a particular environment. The scale of genetic dissimilarity between ecotypes can vary greatly.	Ecotypes must be ecologically distinct enough that they can coexist indefinitely[171]. A mutant within an ecotype can outcompete the other strains in its own ecotype, but not those from a different ecotype[69].
Phylotype	Clade in which all members contain a homologous sequence (marker gene or marker genes, genetic or inter-genic regions) that are distinctively similar.	The threshold level of similarity may be arbitrarily chosen. Not limited to within species.
SNV-type or SNP type	Set of genomes that share a distinctive set of SNVs.[107]	Also used to describe the type of a SNV (for example, the exact switch in nucleotides)
Structural variant	Set of genomes that share distinctive structural variations[172].	Structural variations can be defined as insertions, deletions and inversions greater than 50bp in size[172].
Pathotype	Set of genomes that cause the same disease using the same set of virulence factors[173].	Based on observational data; phenotypic and genotypic. It is not necessarily a clade.
Serotype and serovar	Cells or viruses classified together based on their cell surface antigens, allowing the epidemiologic classification of organisms to the subspecies level[174–176].	Different strains can belong to the same serotype[177]. Certain serotypes are often associated with specific pathotypes[178].
Phagotype (or phage type)	Set of genomes susceptible to a particular bacteriophage and demonstrated by phage typing[179].	Also called 'lysotype'[179].

Figures

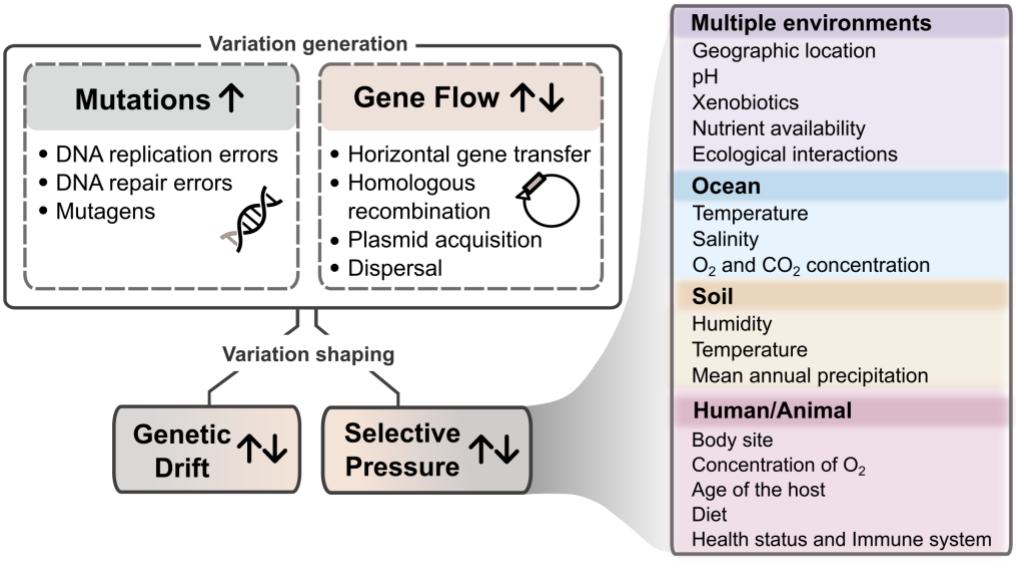

Figure 1. Drivers of variability within bacterial species. Within-species variability is introduced by mutations, which usually increase the amount of variations within a species (up arrow), and gene flow mechanisms, which can increase or decrease the amount of variation within a species. This variability is shaped by genetic drift and selective pressure, which can also increase or decrease the amount of variation. Selective pressures are shaped by many biotic and abiotic factors, some of which are known to drive adaptation in particular habitats more than others.

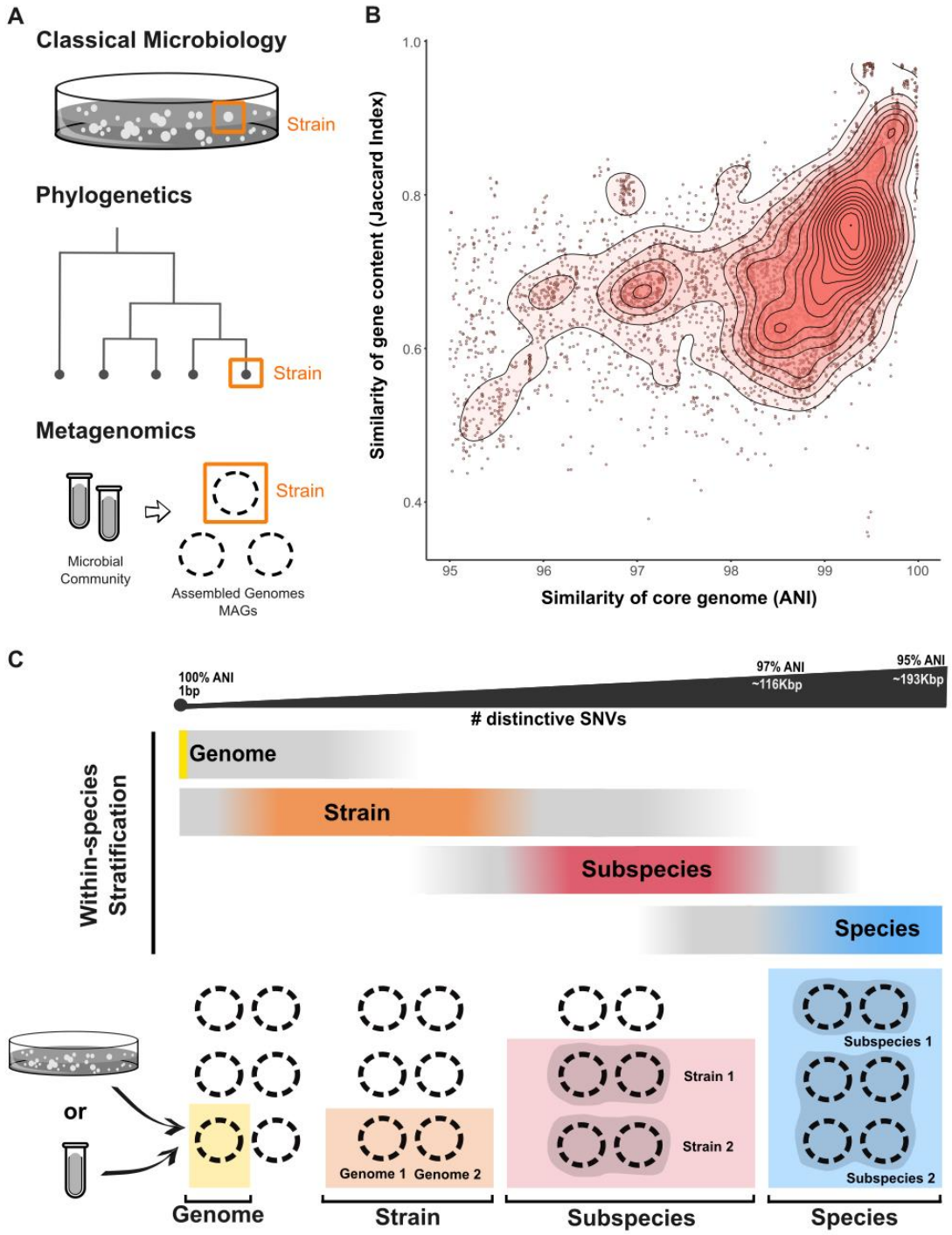

Figure 2. Within-species stratifications (a) Different operational definitions of 'strain', based on the field of investigation: a cultured isolate in classic microbiology, a leaf node in a phylogenetic tree, and a metagenomic assembled genome (MAG) in metagenomics**(b)** Each point is a pairwise-comparison of one isolate genome versus all other conspecific isolate genomes. The data[99] is from 155 bacterial species, each with at least 10 sequenced isolate

genomes. Opacity of red-coloured topographical overlay indicates density of points. The plot shows the relationship between the similarity of the core genome, measured by average nucleotide identity (ANI), versus the similarity of gene content, measured by Jaccard Index. Genomes with higher similarity between their core gene sequences tend to have more genes in common (Spearman correlation R=0.57, p < 2.2e-16). However, high ANI does not necessarily imply highly similar gene content, with many genomes with over 99% core genome ANI having less than 70% of genes in common. Most within-species ANI values are greater than 97%, the few data points below 95% ANI are not shown (83% and 4% of data points, respectively). The data are adapted from Ref[99].**(c)** Spatial distribution of key terminology used to stratify variation within bacterial species, ranging from a single nucleotide variation in the whole genome to the species-level threshold (97% ANI). The coloured portions of the bars reflect the recommended scope of use for each term, and the grey portions indicate the common, often unspecific, scope of use. Broadly speaking, conspecific genomes have identical nucleotides at homologous positions across 97% of their genome (97% ANI), which corresponds to differing on the order of 116,000 SNVs based on an average bacterial genome (3.87Mb[180]). The bottom panel illustrates the hierarchy of these terms, with a species potentially containing multiple subspecies, a subspecies containing multiple strains, and a strain containing multiple (non-identical) genomes. These genomes can be sequenced from cultured isolates or through assembly of a metagenomic sample, creating a MAG which represents the consensus genome of a population of cells.

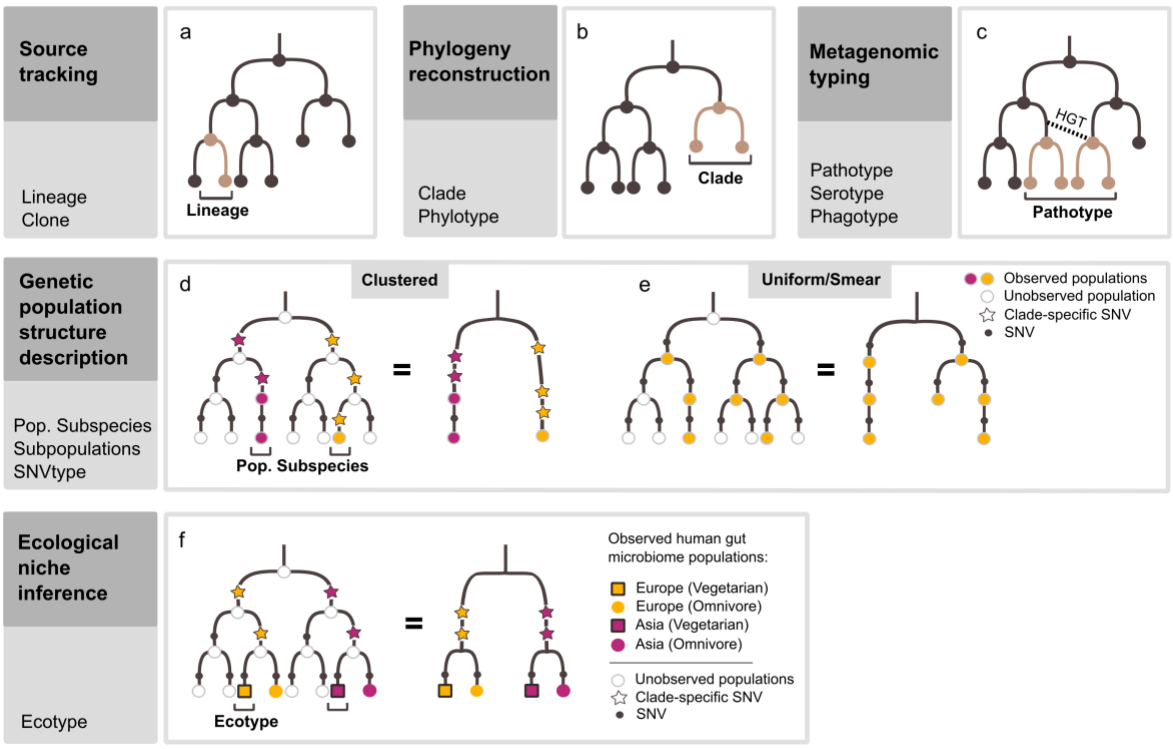

Figure 3. Applications of within-species variation. Five major areas of investigation for within-species-oriented metagenomic data analysis are illustrated, paired with corresponding appropriate terminology. Trees depict the genetic similarity and ancestry of potentially co-existing populations, with nodes representing populations and edges representing genetic differences accumulating from top to bottom. (**a**) Source tracking is concerned with identifying an unbranched path through a tree of ancestors and descendants (a 'lineage', pink edges and nodes). (**b**) Phylogeny reconstruction aims to build a tree which reflects the history of within-species variants based on their genetic similarity. A phylogeny might be cut into complete sub-trees ('clades') which may be called 'phylotypes'. (**c**) Metagenomic typing detects the presence of a previously identified signature of interest within a species. For example, the presence of a gene associated with pathogenicity could be the criteria for detecting a 'pathotype'. This gene may have been transferred between clades via horizontal gene transfer (HGT), so may be at odds with the within-species phylogeny. (**d, e**) The genetic population structure of a species can be described from the distribution of the genetic similarities across observed variants. (**d**) A 'clustered' structure occurs when there is a discontinuity across genetic similarities, enabling clades to be grouped into distinct clusters. Such a non-uniform structure is created by unobserved (extinct or unsampled) intermediate

populations. A hypothetical within-species history with unobserved populations (white nodes) can be simplified (=), showing how unobserved populations can lead to a clustered genetic distribution, which may include distinct population subspecies. As SNVs (black dots) accumulate through this history, some might be specific to a particular set of populations (coloured dots). (**e**) When unobserved intermediate populations are rare or when they are spread widely through a species, the genetic distribution appears uniform or smeared and distinct groups of populations are not seen. (**f**) Ecological niche inference combines population observational data with phenotypic and/or habitat data to identify populations that have adapted to particular niches ('ecotypes'). Adaptive traits might be identified by comparing populations but potential geographic confounds must also be considered.

Box 1. Molecular approaches to characterize variation within species

A wide range of methods are available for studying within-species variation, either based on cultured isolates or directly in microbial communities.

Microbiome-based methods are less established but are not limited to culturable microbiota. Foundational community-fingerprinting methods like DGGE, TRFLP and ARISA[181,182] enabled some species to be studied at high resolution without culturing. Due to their low-throughput and limited resolution, these methods have largely been superseded by genetic sequencing approaches. Despite its origin as a low-resolution method, 16S rRNA gene amplicon analysis methods can sometimes now differentiate within some species using Oligotyping[183,184], amplicon sequence variants (ASV)[185–187] and SNVs in full gene sequences[188]. However, 16S rRNA approaches remain extremely limited in resolution for within-species analysis and can be confounded by multiple, non-identical copies of the 16S rRNA gene per genome[188].

Shotgun metagenomic sequencing provides more information by considering more marker genes or whole genomes. Many tools have been developed to analyse metagenomic data to describe variation within species[19,20]. The major approaches include: SNV-based profiling, either within predefined marker genes[56,90,98,105] or across whole species-reference genomes[103,104,119], overall similarity to strain-reference genomes[92,93], sequence typing[116] and gene content-based profiling[95]. Metagenomic assembled genomes (MAGs) can be recovered

by binning and assembling of co-abundant genes[189]; however, these come with important limitations (**Box 3**).

Non-microbiomic but culture-free methods include microfluidics-based techniques that enable organism-specific enrichment prior to sequencing[190,191]; and single-cell sequencing, which produces single amplified genomes (SAGs)[192]. Culturing is becoming possible for a growing number of bacteria due to methodological advances, such as culturomics, which combines the use of multiple culture conditions with rapid bacterial identification[10].

Non-genomic approaches, such as cryo-electron microscopy-based imaging and transcriptome, proteome- and metabolome-based profiling methods can capture phenotypic differences within species and can be used both separately and in conjunction with genomic approaches. These methods range from well-established, such as serotyping and functional profiling, to more recent and high-throughput, such as thermal proteome profiling[140].

Box 2. Culturing isolates versus metagenomics for analysis of variation within species

Traditionally, investigations below the species level have relied on studying cultured isolates. With the rise of metagenomics, the amount of high-resolution genetic data has increased. Generally, this data is analysed based on variation within specific genetic segments (for example, marker genes) or within genomes recovered through assembly (MAGs) (**Box 1**). Although this enables unprecedented discoveries due to the large scale of data produced, these new methods also have important limitations and introduce new complexity (see the table). Although metagenomics provides important new benefits over studying isolates, the two methods remain complementary[9,193]. To ensure future synergy between the two approaches, isolate genome and metagenomic assembled genome (MAG) data quality must be readily available and comparable, and a common vocabulary should be maintained.

Criteria	Culturing isolates	Metagenomic sequencing
Scope of microorganisms that can be studied below the species level	Must be culturable in isolation but can be low abundance in original sample	Must be abundant or deeply sequenced
Ability to describe multiple species variants within one	Requires multiple rounds of isolation	Can be determined from sequencing data from one

sample (of the same or of different species)	Intractable for low abundance variants	sample, but sufficient sequencing depth required to distinguish from sequencing error
Ability to determine whether genetic variants originate from same organism (genetic linkage)	Possible (as long as variation within isolate colony is low, which is normally the case)	Very difficult or impossible in current typical approach but improvements are possible; for example, long reads, time-series data and Hi-C sequencing[194]
Ability to put species variant in context of community	Limited and work intensive	Implicitly supported, though biases exist[147,161,195]
Ability to describe phenotypic differences between within-species variants	Heterogeneity can be assessed[133] with clinical, environmental and industrial relevance	Limited to description of potential phenotypes
Support for follow-up study	Isolates can be directly experimented on (for example, response to drug exposures)	Extracted DNA can be further tested molecularly (for example, PCR)
Main method for genome recovery	Isolate shotgun genomic sequencing and assembly	Shotgun metagenomic DNA sequencing followed by assembly (MAG; **Box 1**)
Quality of the recovered genomes	Often remain at draft level but usually are high quality with little contamination	May have higher error rates and be chimeric, contaminated and incomplete[192,196]
Quality assessment of the recovered genomes	Provided by central repositories, with various guidelines developed (see, for example, Ref.[42])	Routinely assessed but *ad hoc* by authors Recommendations are emerging[192]
Determining presence or absence of gene in the recovered genome	Usually simple and correct	Difficult to be certain
Expected impact of long read sequencing	Longer contigs, less challenged by repetitive regions	Better genomes for the most abundant organisms, low abundance fraction still hard to access

Box 3. Challenges in studying variation within species in metagenomics

Investigations of variation within species in microbial communities are faced with study-design, technical and methodological challenges. A main study-design challenge is the

’unobserved variation’ paradigm: you do not see what you do not sample. If low variability is seen within a species, it is difficult to prove that it is not due to under-sampling or sampling bias. This bias can be temporal (for example, due to strain turnover or extinctions) or spatial (for example, due to proximate sampling areas, such as soil or skin, harbouring substantially different infraspecific profiles). Shallow sequencing depth also biases against observing low abundance within-species variants. These biases are mitigated by the increasing number of deeply sequenced metagenomic samples. However, integration of these samples across studies is still faced with technical challenges well-known in metagenomics[195,197–199].

Although undoubtedly useful for investigating unknown and under-represented species, metagenomic assembled genomes (MAGs) have important limitations. MAGs are population consensus genomes, thus, loci may be polyallelic and unlinked[164,196]. When compared to isolate genomes, MAGs often have low assembly quality, are less complete and are more likely chimeric[103,164,192,196,200]. For these reasons, and due to the difficulty in detecting chimeras below the species level, MAGs should not be considered equivalent to genomes sequenced from isolates[196]. The use of the term ’complete MAG’ (CMAG) should be adopted only for MAGs that are analogous to complete isolate genomes, which are usually a single circular contig with no gaps.

To avoid confusing isolate genomes and MAGs, the growing practice of uploading MAGs to public genome databases[196] should be discouraged and the phrase ’genome-resolved metagenomics’ should not be used for MAG studies that do not directly assess heterogeneity within MAGs. Single-cell sequencing approaches provide a promising alternative to MAGs for recovering genomes from metagenomes, but are limited by high cost, low throughput, potential contamination and quality issues due to using a single molecule of DNA[201].

Continued technical advances, decreasing sequencing costs, and increasing integration of complementary methodologies will be necessary to counteract these challenges in data generation and integration.

Glossary

Metagenomics

The study of all genomes present in a sample from a microbial community. Often performed as shotgun metagenomics, in which extracted DNA is fragmented before sequencing.

Population

A set of individuals that occupy a particular spatial area

Mutator allele

Genetic variation (allele) that results in an increased mutation rate.

Horizontal gene transfer (HGT)

The movement of genetic information between organisms. This is in contrast to vertical gene transfer from parent to offspring.

Homologous recombination

Type of genetic recombination in which genetic material is exchanged between two similar or identical regions of DNA.

Conspecific

Belonging to the same species. For example, conspecific strains are strains that belong to the same species.

Genetic drift

Change of allele frequencies in a population caused by stochastic factors

Marker genes

In microbiome context: genes or genetic segments whose presence or specific DNA sequence is distinctive of a category of interest, such as a species or clade.

Selective sweep

Reduction of the genetic variation in a population due to selection acting on novel mutations or existing alleles.

Hard selective sweep

One beneficial allele at a locus replaces most other alleles in the population.

Soft selective sweep

Multiple beneficial alleles at a locus gain prevalence, replacing standing genetic variation in the population.

Metagenomic assembled genome (MAG)

A genome sequence recovered from metagenomic data, usually fragmented, and potentially incomplete or contaminated. Typically, shotgun metagenomic sequencing produces short DNA sequences that are then assembled and binned into 'genomes' using k-mer frequencies and abundance information.

Type strain

A living culture that serves as a fixed reference point for the assignment of bacterial and archaeal names. It is descended from the original isolate used in a species' description and shares all of its relevant phenotypic and genotypic properties.

Microbiomics

The study of microbial communities (microbiomes) using one or more '-omic' approaches (e.g. genomics, transcriptomics, proteomics, etc.)

Infraspecific

Below species level, that is, at a higher resolution than 'species'.

Polyphyletic

Describes a group of organisms that do not share an immediate common ancestor. Not a clade.

Guilds

A guild is a group of species that use the same type of resources in a similar way. Originally defined as a group of species (Root, 1967) but concept could be applied to strains or subspecies.

Genome-wide sweep

Alleles at the locus under selection cause other linked loci (for example, genome, plasmid) to gain or lose abundance across the population. Also known as a 'broad' sweep.

Gene-specific sweep

Only alleles at the locus under selection gain or lose abundance across the population. Also known as a 'narrow' or 'locus-specific' sweep.