# Peering into the Genetic Makeup of Natural Microbial Populations Using Metagenomics


**Vincent J. Denef**

**Abstract** This chapter focuses on how metagenomic data are applied to examine the genomic heterogeneity of natural microbial populations. It highlights the opportunities and challenges inherent to the approach and describes recently developed methods to maximally leverage the potential of these datasets while tackling some of the challenges. We describe how performing population genomic analyses using metagenomic data allows (1) resolution of ecologically and genetically cohesive populations in the environment, (2) tracking of evolutionary processes within them, and (3) application of metatranscriptomic and metaproteomic analyses to determine the in situ physiology of distinct populations. While challenges remain that are inherent to the approach, the current wave of new bioinformatic tools is starting to realize the theoretical potential of metagenomics to peer into the spatiotemporal dynamics of the genetic structure of natural populations.

**Keywords** Bacteria · Bioinformatics · Gene content variation · Metagenomics · Natural populations · Recombination · Selection · Sequence variation · Strain-resolved

## 1 Introduction

### *1.1 Scope of This Chapter*

This chapter explores the advances that have been made in the area of population genomics through studies that make use of metagenomic data. It covers methodological advances and challenges and biological insights that have been gathered. Metagenomics [also called environmental or community genomics (Handelsman

V. J. Denef (✉)
Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, MI, USA
e-mail: vdenef@umich.edu

Martin F. Polz and Om P. Rajora (eds.), *Population Genomics: Microorganisms*, Population Genomics [Om P. Rajora (Editor-in-Chief)],
https://doi.org/10.1007/13836_2018_14,

2004; Tyson et al. 2004)] is the analysis of data produced through randomly sequencing fragments from a DNA pool extracted from environmental samples (Fig. 1). In general, metagenomics finds its application in the study of the composition and functional potential of microbial communities in their native environment. However, a growing number of studies are leveraging these datasets to gain unprecedented insights into the genetic composition of natural populations, i.e., groups of individuals belonging to the same species that co-occur in space and time. In the context of metagenomics, the term population genomics was initially—and continues to be—used as a synonym for metagenomics, a possible cause of confusion. This has been particularly the case when referring to the reconstruction of a consensus genome of a population through curated assembly of metagenomic data (DeLong 2004, 2005; Handelsman 2004; Tyson et al. 2004). In the context of this book chapter, the strictest definition of metagenomic-based population genomics refers to the analysis of genome-wide heterogeneity existing between individuals belonging to the same species/ecotype (Whitaker and Banfield 2006).

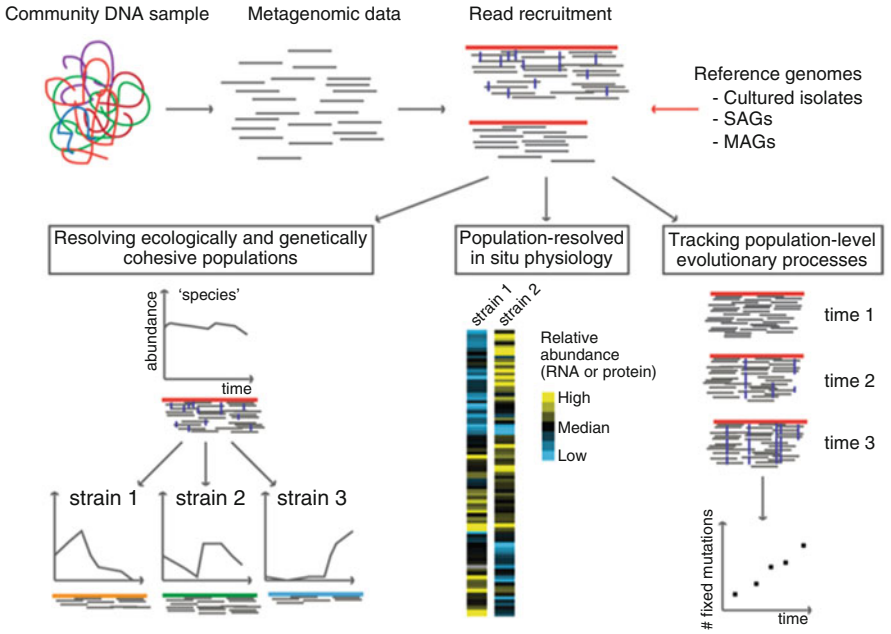


**Fig. 1** Overview of applications of metagenomic-based population genomics discussed in this chapter. After extracting, fragmenting, and generating sequencing reads to create a metagenomic dataset, reads are typically aligned to a reference sequence, obtained from microbial isolates, single cells (SAGs), or assembled from metagenomic data (MAGs). These data can be used to (left) resolve ecologically distinct populations by identifying genetically similar reads originating from distinct strains, (middle) determine in situ gene expression (using metatranscriptomic or metaproteomic data), or (right) track evolutionary processes, for example, by identifying polymorphic sites where specific nucleotides rise to fixation over time

While more theoretically motivated definition of populations is the focus of another chapter in this book (Shapiro 2017), it is important to highlight how population genomics using metagenomic data has contributed to our efforts to recognize and delineate ecologically and evolutionary cohesive populations. Most prominent is the demonstration of sequence-discrete populations in environmental samples [Fig. 2; sensu (Caro-Quintero and Konstantinidis 2012)]. Conceptually, a sequence-discrete population was defined by Caro-Quintero and Konstantinidis as "the natural entity present in a community/sample that comprises genotypes, which are clearly distinguishable from their closest co-occurring relatives (if any) based on their high genetic relatedness and comparable relative abundance *in situ*." Technically, such genotypic clusters are identified by comparing the nucleotide identity of the short reads gathered in a metagenomic survey to a reference genome by means of a process called read recruitment. Whereas similar observations of sequence-discrete populations based on a single or multiple marker genes sequenced from bacterial isolates had been made before (Hanage et al. 2006; Hunt et al. 2008; Rocap et al. 2003), the advent of metagenomic methodology allowed for a genome-wide assessment of genetic relatedness of randomly sampled cells present in environmental samples. These insights from isolate and metagenomic studies have helped move forward the discussion regarding the existence of microbial species and specifically how to define a microbial population. While genotypic variation within a defined species can be large at a regional or global scale, thus complicating our ability to define clear species boundaries, it is important to stress the inherent property of a population to contain only individuals that are occurring in the same place at the same time, i.e., that they are sympatric (Shapiro and Polz 2014; Cordero and Polz

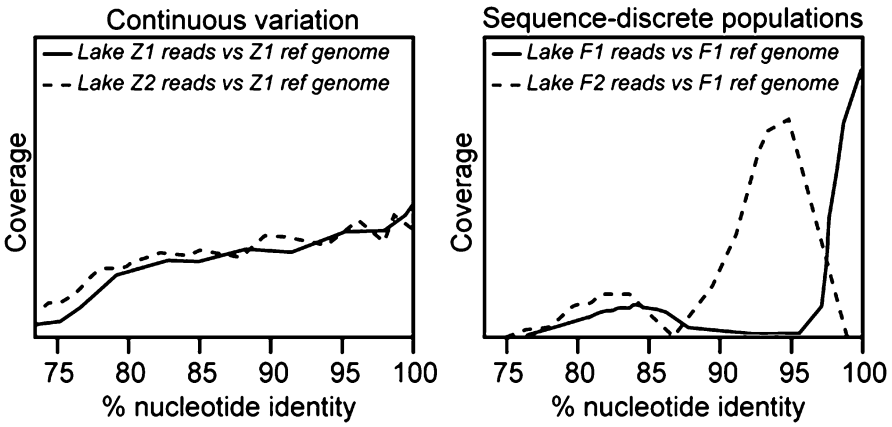


**Fig. 2** Identification of populations using metagenomic data. Comparisons of sequences generated from community DNA samples from two lakes to a reference genome of a bacterial isolate from one of these two lakes. If discrete populations would not exist, the patterns on the left could be obtained, while if ecologically and evolutionary cohesive populations that are distinct from the reference population exist, the pattern on the right would be expected. The patterns on the right are the most commonly observed [e.g., (Bendall et al. 2016, Caro-Quintero and Konstantinidis 2012)]

2014). While implied by the definition of a population, early challenges to the idea that discrete populations exist based on sequenced isolates from disparate locations did not respect this condition of sympatry [e.g., (Welch et al. 2002)]. Further discussion of insights into the microbial species concept derived from bacterial population genomics is covered in other chapters (Shapiro and Polz 2015; Shapiro 2017) and will not be discussed in detail here.

## *1.2 Approaches Included in This Chapter*

Examining population-level variability within natural populations requires the availability of a reference sequence in most metagenomic approaches to population genomics. There are multiple ways to obtain reference genomes: (1) from isolates, preferably originating from the same environment/sample the metagenomic data are derived from; (2) from genomes assembled from metagenomic data (metagenomic assembled genome, MAG), which tends to be a composite sequence not representative of a single cell in the population; or (3) from a single cell genomic dataset (single amplified genome, SAG). While our ability to generate MAGs was initially limited to low complexity communities (Tyson et al. 2004), we can now reconstruct 100 s to 1000 s of genomes from metagenomic datasets. Few of these MAGs are complete, and although contamination from other populations cannot be completely excluded, obtaining >90% completeness with limited (<5%) contamination is commonly achieved (Anantharaman et al. 2016; Delmont et al. 2017; Parks et al. 2017). Tools to refine sequence bins and estimate their completeness and purity are continuously being developed (Broeksema et al. 2017; Eren et al. 2015; Parks et al. 2015) and will continue to improve genome reconstruction and bin refinement, so as to provide a sounder basis for downstream population genomic analyses. More uncertainty will remain in genomes assembled from metagenomes compared to those from isolates.

Independent of the approach used to generate the reference genome, the central tenet of the analysis is the comparison of randomly sampled and sequenced DNA fragments with this reference genome to assess sequence content and compositional variation of populations within and between environmental samples. A variety of approaches to generate metagenomic DNA fragments can be applied. Most straightforward is to randomly generate sequences from DNA extracted from environmental samples. Other approaches first reduce the diversity of the community, e.g., by passing the sample through a series of filters with decreasing pore sizes (Baker et al. 2010) or through (in situ) enrichments (Delmont et al. 2015). This allows for the enrichment of specific populations of interest, therefore increasing sequencing depth for population genomic analyses. Finally, instead of focusing on the entire genome, one can target a series of sites across the genome. An approach recently implemented, which may see much broader application, is to extract multi-locus sequencing typing (MLST) genes from complex metagenomic datasets using reference sequences (Berry et al. 2017; Zolfo et al. 2017).

Although the work discussed in this chapter relates to studies that use random sequencing of DNA extracted from complex natural communities, occasionally single cell genomics is categorized as a metagenomic approach. Single cell sequencing approaches sort a single cell from an environmental sample using dilution, flow cytometry, or microfluidic approaches and subsequently perform DNA amplification and genome sequencing (Blainey 2013). Genome sequencing analyses of single cells representing the same naturally occurring population is similar to metagenomics, as both eliminate culturing biases. Other than that aspect, population genomics based on single cell sequencing [e.g., (Kashtan et al. 2017; Malmstrom et al. 2013; Zaremba-Niedzwiedzka et al. 2013)] is conceptually and methodologically similar to the analysis of representative isolate genomes from natural populations [e.g., (Hunt et al. 2008)], discussed elsewhere in this book. Testament to the power of this approach, Kashtan and colleagues used single cell genomic data of *Prochlorococcus* cells from ocean water to demonstrate the existence of hundreds of co-occurring populations. These populations were shown to differ from each other in terms of genome content, such as the presence of small genomic islands that most likely conferred predation resistance and phage recognition), as well as genome-wide sequence composition (Kashtan et al. 2014). These insights were gained by characterizing >1,000 cells by sequencing of the rRNA ITS region, while a subset of 69 cells was sequenced to >70% estimated genome completeness. As mentioned above, single cell genome sequencing is also commonly used to generate a reference sequence, after which sequencing reads from metagenomic surveys from the same or different environments can be aligned to this genome to evaluate population-level heterogeneity, which is of relevance to this chapter [e.g., (Thrash et al. 2014)].

## 2 Opportunities and Challenges

### *2.1 Opportunities*

The advantage of metagenomic approaches compared to single isolate approaches is the ability to sample a very high number of individuals without culturing bias. While the number of individual cells that can be analyzed is rapidly increasing for single cell approaches, the number of cells sampled and typical reconstructed genome completeness remains higher in metagenomic approaches. This offers the unprecedented ability to peer deeply into the genetic structure of natural populations and has revealed the extraordinary genetic diversity that exists among groups of closely related microbes. The extent of this diversity and the correlation between the abundance of genetic subclusters with distinct environmental conditions can result in the division of previously named taxonomic units into ecologically distinct populations (Bhaya et al. 2007; Denef et al. 2010a). A hallmark of such newly defined populations is extensive diversity in gene content and sequence composition (Denef et al. 2010a; Simmons et al. 2008). Initially, there were doubts that environments beyond reduced complexity systems such as the acid mine drainage system, in

which pioneering studies of population genomics using metagenomics were conducted, could be tackled (Denef et al. 2010b). However, recent work has expanded the approach to systems ranging from the human microbiome to aquatic environments (Bendall et al. 2016; Nayfach et al. 2016; Olm et al. 2017). Importantly, the use of metagenomic approaches allows us to access 10 s–1000 s of populations at the same time (Anantharaman et al. 2016; Parks et al. 2017). The tremendous growth of publicly available metagenomic datasets, as well as reference genomes from microbial isolates or single cell genomics projects, is another major opportunity to tackle new population genomic questions without additional sequencing efforts. This was demonstrated in recent studies that have leveraged thousands of human microbiome metagenomic datasets to uncover strain-level dynamics and infer mode of transmission and biogeographical patterns among hundreds of bacterial populations simultaneously (Nayfach et al. 2016; Truong et al. 2017).

## *2.2 Challenges*

However, many challenges remain (Table 1). The first and possibly most important major challenge is the lack of linkage between variant sites (nucleotide substitutions, insertions and deletions, rearrangements), i.e., using metagenomic data we are unable to determine which alleles across the genome are present in one lineage versus another. As a consequence, most environmental population genomic approaches have relied on isolate or single cell-derived sequences (Kashtan et al. 2014; Krause and Whitaker 2015; Shapiro and Polz 2014). Multiple factors can be responsible for lack of linkage in metagenomic data: (1) the number of variant loci across the genome is typically too low compared to sequencing read length (generally 100–150 nucleotides) or sequencing library fragment size (up to several hundred nt) to enable linkage across more than a few hundred nucleotides, and (2) in metagenomic datasets, each sequencing read typically originates from a different individual. As a result, we are limited to identifying which sites are polymorphic in the population or which sites are divergent between coexisting closely related populations. Determining which mutations occur across the genomes of a single lineage remains possible only by using isolate or single cell genomic data unless population structure is very simple (Denef and Banfield 2012), although new approaches based on statistical inferences may change this (e.g., DESMAN, see below). Considering these challenges due to read length of the most commonly used next-generation sequencing platforms, it is thus not surprising that some of the most thorough metagenomic population genomic work has been carried out using longer sequences, such as those obtained by Sanger sequencing. These studies were largely successful because long sequence reads and library insert size enable linkage across several kilobases at a time (Allen et al. 2007; Eppley et al. 2007; Konstantinidis and DeLong 2008; Simmons et al. 2008). Technological innovation is ongoing, and newer sequencing platforms (e.g., PacBio, Oxford Nanopore) may resolve the read length issues, as long as they continue to offer high sampling depth and limited

**Table 1** Fundamental challenges for metagenomic-based population genomic analyses

| Challenge | Approaches to address challenge | Remaining issues | Example studies |
|---|---|---|---|
| Linking SNPs co-occurring in the same individual | Link by relative abundance | Only works for low within-population genetic diversity or requires high number of samples | Denef and Banfield (2012); Quince et al. (2017) |
| | Focus on overall patterns of polymorphisms across genome | Limited to broad interpretations regarding genome-wide vs gene-specific selective sweeps | Bendall et al. (2016) |
| | Long-read sequencing technologies | High error rates, or lower sequence coverage, or not broadly available | Sharon et al. (2015) |
| Differentiating SNPs vs errors | Tool for identifying true sequence variants from sequencing error (Varcap) | Only helps resolve SNPs present in >2% of population | Zojer et al. (2017) |
| | Tools for removing error-based bias in population genetic parameter calculations | Platform-specific, unclear if it removes sequence-library-dependent bias | Johnson and Slatkin (2006); Johnson and Slatkin (2008); Johnson and Slatkin (2009); Nielsen et al. (2011) |
| Obtaining sufficient sequence coverage | Physical- or affinity-based enrichment of target populations | Affinity-based methods often technically challenging, physical methods restricted to populations with outlier cell size | Baker et al. (2006); Hatzenpichler et al. (2016); Pernthaler et al. (2008) |
| Tracking gene gain/loss | Reference-free sequencing read dataset comparisons | Untested for population genomic analyses | Nijkamp et al. (2013) |
| | Sample-by-sample genome reconstruction | Restricted to populations that are abundant across time series | Bendall et al. (2016) |

sequencing errors (see below), which is typically not yet the case [but see (Sharon et al. 2015)]. Alternatively, methods that use cross-linking of DNA within the cell may allow the connection of physically linked variants (Marbouty et al. 2014), but this method has not been applied to population genomic studies.

The second major challenge is the issue of sequencing error, which results in false positive polymorphic sites. These errors usually occur as low-frequency "mutations" that are often observed only once (Schirmer et al. 2016). As several population genetic parameters require knowledge of all polymorphic sites, including those occurring at low frequency, errors will bias their metagenomics-based estimates. Case in point is Watterson's theta (Watterson 1975), which estimates the genetic diversity present in a population based on which we can estimate mutation rates and/or effective population sizes and which requires even the knowledge of the frequency of singleton variant sites. As new sequencing platforms have emerged,

each with their specific error spectrum, a variety of tools have been developed to differentiate between true variants and sequencing errors. Some tools assign confidence levels to observed variants, with the goal of reducing false positive variants [e.g., VarCap (Zojer et al. 2017)] focuses on improving reliability of identifying variants >2% of the population), or to remove bias in population genetic parameter estimates (Johnson and Slatkin 2006, 2008, 2009; Nielsen et al. 2011). In addition to sequencing platform-specific error profiles (Schirmer et al. 2016), a complication arises from the observation that errors can be sequencing library preparation protocol dependent. This issue was highlighted recently in a reanalysis of the preterm fecal microbiome metagenomic data that is frequently used as a benchmark dataset for new bioinformatic tools (Sharon et al. 2013). It could be shown that the day-by-day alternation between two SNP patterns in a bacterial population was caused by different library preparation methods used on even and odd days [http://merenlab.org/2016/12/14/coverage-variation/ comment on (Eren et al. 2015) based on data from (Sharon et al. 2013)].

The third challenge is obtaining sufficient sampling depth, which is the number of sequences that cover a particular region of the genome. Sufficient sequence coverage is needed to accurately estimate allele frequencies, and as such most analyses are currently limited to the most abundant populations in environmental samples. Yet, we have progressed far beyond what the research community envisioned just a few years ago with respect to the number of near-complete genomes that we can reconstruct from environmental samples. Such genomes can subsequently be used to examine the genetic structure of the corresponding populations. In part, this is due to the development of physical (e.g., size-selective filtration)- or affinity (e.g., based on use of fluorescent in situ hybridization and cell sorting)-based methods (Baker et al. 2006; Hatzenpichler et al. 2016; Pernthaler et al. 2008) through which populations of interest (e.g., based on taxonomic identity or metabolic activity) can be enriched, facilitating population genomic analysis from the corresponding metagenomic data (Deng et al. 2014).

Finally, the identification of gene content differences within a population can be challenging when using metagenomic data. Unless extensive manual curation of an assembly is performed [e.g., (Simmons et al. 2008)], genomic regions (islands) carried only by low-abundance subpopulations (i.e., part of a population's "flexible genome") will generally not be binned in the consensus genome of the population of interest (i.e., the "core genome"). This is because these genomic regions (a) diverge in their k-mer (specific stretches of nucleotides, e.g., tetramers ATGC, AATG, etc.) composition [used in most binning applications that seek to group fragments of contiguous sequence (contigs) originating from the same genome such as VizBin (Laczny et al. 2015), CONCOCT (Alneberg et al. 2014), and TETRA-ESOM (Dick et al. 2009)] compared to core genome contigs, (b) have differential coverage patterns that diverge from the core genome of the population [also commonly used in binning applications (e.g., CONCOCT (Alneberg et al. 2014), GroopM (Imelfort et al. 2014), metagenome (Albertsen et al. 2013))], and/or (c) will often fail to assemble into large enough contigs to allow accurate binning due to the low abundance of the subpopulations they derive from. Similarly, when using

metagenomic sequences to identify variants across datasets by mapping sequence reads to reference genomes, we can only track changes in frequency among regions shared by these populations and cannot identify the addition of new genomic regions due to horizontal gene transfer [see (DeLong 2012) in commentary on (Denef and Banfield 2012)]. Yet, we know that such differences constitute a significant fraction of population-level genomic heterogeneity. Evidence regarding the physiological importance of these unique regions is mixed (Denef et al. 2010a; Frias-Lopez et al. 2008; Gogarten and Townsend 2005; Kuo and Ochman 2009; Thompson et al. 2011; Hehemann et al. 2016). Nonetheless, it is important to try to include these regions in metagenomic-enabled population genomic analyses as gene frequencies at either intermediate or low levels result from frequency-dependent selective pressures by social and ecological interactions and thus suggests adaptive roles for flexible genome content (Coleman et al. 2006; Cordero et al. 2012; Cordero and Polz 2014; Kashtan et al. 2014; Rodriguez-Valera et al. 2016).

## 3 Current Applications

We present a series of recently developed tools to facilitate population genomic analyses using metagenomic data and explore three types of applications of these methods (Fig. 1; Table 2). First, we provide an overview of how these methods are being used to resolve ecologically and genetically distinct populations that would previously have been considered as a single operational taxonomic unit (OTU). Most commonly OTUs are defined based on 16S rRNA gene sequence identity, but these can similarly be defined based on multiple housekeeping genes or complete genomes. Second, we show how these approaches can be used to infer the physiology of distinct populations. Third, we summarize applications of these methods to gather insights into evolutionary processes occurring in natural microbial populations.

### *3.1 Methods*

Read mapping to a reference sequence is a key step in most population genomic approaches. Over time a wide array of read alignment tools have become available, each with their own user-specified tunable parameters. Naturally, this flexibility may affect our ability to accurately perform population genomic analyses. In a recent comparative analysis, popular tools such as bwa (Li and Durbin 2010) and bowtie2 (Langmead and Salzberg 2012) resulted in similar and more accurate results than some other tools when all were run using default parameter settings (http://merenlab.org/2015/06/23/comparing-different-mapping-software/). In more recent years, reference-free methods have been developed that avoid some drawbacks of the reference-based approach, particularly the inability to detect parts of the population's

**Table 2** Goals, approaches, and challenges for metagenomic-based population genomic analyses

| Goal | | Approaches | Challenges | Example studies |
|---|---|---|---|---|
| Resolving ecologically and genetically cohesive populations | Identifying sequence-discrete populations | Read recruitment (e.g., bowtie2, bwa) + custom scripts for data plotting | Determine relevant sampling scales to capture sympatric individuals (e.g., bulk water vs size-fractionated samples) | Caro-Quintero and Konstantinidis (2012) |
| | Distinguish diverging within-species ecological dynamics | Growing suite of automated tools such as Constrains, MetaMLST, MIDAS, PanPhlAn, StrainPhlAn, DESMAN, and LSA | Database dependency of many tools limits us to species with extensive reference genome availability | Luo et al. (2015); Zolfo et al. (2017); Nayfach et al. (2016); Asnicar et al. (2017); Ward et al. (2016); Quince et al. (2017); Cleary et al. (2015) |
| | Identify strain-specific gene content and SNPs | | Most approaches need a large number of samples to be effective | |
| Determining physiology of ecologically and genetically cohesive populations | Identify in situ differences in gene expression between co-occurring strains | Custom scripts/manual as well as automated tools to resolve metatranscriptomic or metaproteomic data (e.g., PanPhlAn) | Relationship between expression levels and process rates rarely known | Wilmes et al. (2008); Denef et al. (2010a); Brooks et al. (2015); Asnicar et al. (2017) |
| | Estimate in situ growth rates | iRep, based on metagenomic coverage patterns | Not benchmarked against measured growth rates thus far | Olm et al. (2017) |
| Tracking evolutionary processes within ecologically and genetically cohesive populations | Homologous recombination vs mutation | Manual tools (e.g., Strainer) to visually identify and quantify recombination sites or automated tools to determine recombination vs mutation rates | Manual work is low throughput | Eppley et al. (2007); Johnson and Slatkin (2009) |
| | Gene gain/loss | Custom scripts to determine gene content differences between MAGs representing same population across time series samples | Can we discriminate gene gain/loss in population vs strain replacement? | Bendall et al. (2016) |
| | Natural selection | Custom scripts/manual approach to determine mutation rates and/or gene-specific vs genome-wide selective sweeps | Challenging in "open" systems | Denef and Banfield (2012); Roux et al. (2014); Bendall et al. (2016) |

flexible genome. One such tool is able to detect gene frequency patterns across samples of regions of the genome that are not represented in reference sequences [e.g., MARYGOLD (Nijkamp et al. 2013)]. Historically, population genomic analysis of metagenomic data relied on manual analysis, either through existing assembly visualization/curation software (Morowitz et al. 2011; Simmons et al. 2008), through generic graphing software (e.g., excel, R) to visualize the distribution of sequence similarities among reads mapping to a population's contigs (Fig. 2; [Bendall et al. 2016; Caro-Quintero and Konstantinidis 2012; Oh et al. 2011)], or through software developed specifically for the resolution of bacterial strains in metagenomic data [e.g., Strainer (Eppley et al. 2007)]. While these approaches worked, a drawback of these methods is the labor-intensiveness and difficulty to reproduce similar results by independent users with different levels of expertise.

More recently, a variety of tools have been developed to (1) remove bias due to sequencing error [VarCap (Zojer et al. 2017)], (2) extract population genomic metrics (e.g., Watterson's theta) from next-generation sequencing data (Haubold et al. 2010; Johnson and Slatkin 2006), (3) visualize SNP patterns across sample series in assembled contigs [e.g., using Anvi'o (Eren et al. 2015)], and (4) resolve closely related strains from metagenomic datasets [e.g., ConStrains (Luo et al. 2015), MIDAS (Nayfach et al. 2016), DESMAN (Quince et al. 2017), StrainPhlAn (Truong et al. 2017), PanPhlAn (Scholz et al. 2016), and LSA (Cleary et al. 2015)]. The development of the latter set of tools is particularly exciting, as it promises to greatly facilitate the resolution of strain dynamics and the coupling of gene content and sequence composition data with dynamics in population abundance across environmental or temporal gradients as has been performed manually previously (Denef et al. 2010a; Morowitz et al. 2011).

Most of the strain resolution tools rely heavily on whole genome reference databases, which are reasonably representative for some microbial systems such as the human microbiome, but much less so for other systems such as terrestrial and aquatic biomes. The reliance on reference genomes limits the ability for strain resolution in these other environments at this point (Nayfach et al. 2016). All of the reference-based tools are able to analyze thousands of metagenomic datasets at the same time while extracting strain dynamics for many species at the same time, e.g., 135 in the case of the study by Truong et al. (2017). Based on their own benchmark study, StrainPhlAn appears to reduce the per-nucleotide nucleotide variant identification error to less than 0.1%, granting more accurate strain identification than tools such as MIDAS and ConStrains. PanPhlAn is similar in approach to StrainPhlAn, but is focused on identifying strain-specific gene content, rather than nucleotide substitutions (Scholz et al. 2016).

In contrast, strain resolution tools such as DESMAN and LSA take reference sequence-independent approaches and, in the case of LSA, even an assembly-independent approach. DESMAN identifies strains, including both genotype-specific nucleotide substitutions and gene content variation, from metagenomic data generated from a sample collection. After validating their approach with a mock dataset, they applied their method to examine abundance patterns of different strains within a large set of ocean metagenomic data (TARA Oceans). While

DESMAN allows the identification of novel strains, it is highly dependent on the quality of the assembly and binning steps and requires a relatively large number of samples to be effective (Quince et al. 2017). Many researchers currently rely on automatic binning approaches to generate their metagenomic sequence bins, but these can be highly inaccurate, depending on community composition including the extent of co-occurring closely related populations and the extent of community turnover in the temporal or spatial sample series. While DESMAN has the ability to further resolve multi-strain bins, careful manual curation, aided by tools such as Anvi'o (Delmont et al. 2017) or ICoVeR (Broeksema et al. 2017), may be necessary for downstream population genomic analyses. The second assembly-independent approach discussed here, latent genome analysis (LSA; Cleary et al. 2015), separates sequencing reads prior to assembly by calculating unobserved variables call "eigengenomes" that reflect covariance in k-mer abundances across a sample series. This method allowed for the separation and downstream assembly of specific genomic regions of strains sharing less than 99.5% average nucleotide identity, while regions of the genome highly conserved between strains were grouped together as sequencing reads from which conserved core genome could be assembled.

Beyond whole genome approaches, several approaches have been developed to extract population-specific sequences for a set of core genes. The concept of multi-locus sequence typing (MLST) (Maiden et al. 1998) used for population genetic analysis of isolates has been implemented in metagenomic data analysis either through a series of custom bioinformatic scripts (Berry et al. 2017) or through more streamlined packages such as MetaMLST (Zolfo et al. 2017), ConStrains (Luo et al. 2015), and MetaPhlAn2 (Truong et al. 2015). Finally, to resolve true sequence variants from sequencing errors, tools such as oligotyping (Eren et al. 2013) can be applied to sequence reads covering marker genes, though we are not aware of applications to metagenomic data thus far.

## *3.2 Resolving Ecologically and Genetically Cohesive Populations*

The motivation to develop these new metagenomic tools originated from the realization that studies using single marker genes clustered at a fixed identity level (i.e., OTUs) likely miss key community dynamics since multiple ecologically distinct populations were clustered together in a single OTU (Acinas et al. 2004; Denef et al. 2010a; Eckburg et al. 2005; Fraser et al. 2009; Fuhrman and Campbell 1998; Giovannoni et al. 1990; Hahn et al. 2016; Hunt et al. 2008; Larkin and Martiny 2017; Morowitz et al. 2011; Rocap et al. 2003; Shapiro et al. 2012; Shapiro and Polz 2014; Sharon et al. 2013; Wilmes et al. 2008). The ability to resolve strain-level

differences in microbial communities and detect the dynamics of highly related genotypes will likely lead to rapid advances in our ability to study microbial ecology at the appropriate resolution. We present here some of the most recent examples of how streamlined strain-resolved analyses are leading to previously unrecognized ecological patterns.

Using MIDAS, researchers were able to identify strains in metagenomic datasets and this revealed dynamics that could not be observed at a higher taxonomic level (e.g., species) (Nayfach et al. 2016). Conceptually, the finding that important ecological dynamics are masked by clustering distinct populations into higher taxonomic levels is similar to previous findings. Particularly, a study by Rodriguez-Brito and coauthors showed that hidden underneath the observed stability at coarser genetic resolution ("species" level) were strongly fluctuating abundances of ecologically distinct "strains" grouped at the species level (Rodriguez-Brito et al. 2010). The study by Nayfach and coauthors revealed that mothers pass on a large percentage of bacterial allele variants to their children in the early days after birth. In the subsequent postnatal months, even as the number of species shared between mother and child increases, the strain composition gradually diverges (Fig. 3a, b), indicating increasing importance of colonization from other sources (Nayfach et al. 2016). These findings were confirmed in a similar study using PanPhlAn and StrainPhlAn (Asnicar et al. 2017). At a larger spatial scale, links between geographic distance and strain correspondence have been found in human populations using StrainPhlAn as well and indicate limited overlap in strains between geographically distinct populations (Truong et al. 2017).

Several of the recently developed methods allow us to pinpoint the specific gene content and SNP variation that differentiates closely related but ecologically distinct populations from each other to attempt to explain their distinct population dynamics. For example, resolution of strains and identifying strain-specific gene content has allowed for the identification of specific strains involved in diseases where traditional approaches failed to do so. Using PanPhlAn, Ward and coauthors identified strain-specific gene content of *Escherichia coli* using 166 infant microbiomes and identified strains associated with infant risk for necrotizing enterocolitis to be enriched in genes involved in iron acquisition and specific energy and amino acid metabolism functions (Ward et al. 2016). In another study, an analysis of regional strain-level variability identified regionally distinct horizontally transferred genes, in large part glycosyl transferase family proteins likely reflecting dietary differences at both large and small spatial scales (Brito et al. 2016) (Fig. 3c). While these studies did not aim to resolve co-occurring closely related populations, the same approach could be applied to identify genes differentiating sympatric populations.

The field of epidemiology is also embracing metagenomic tools to better understand disease outbreaks. As MLST is a common method used in epidemiological studies using isolates, tools adapted to metagenomic data, such as MetaMLST, have been used to identify strains in disease outbreaks (Zolfo et al. 2017). In addition, the large number of reference sequences available for pathogenic bacteria in

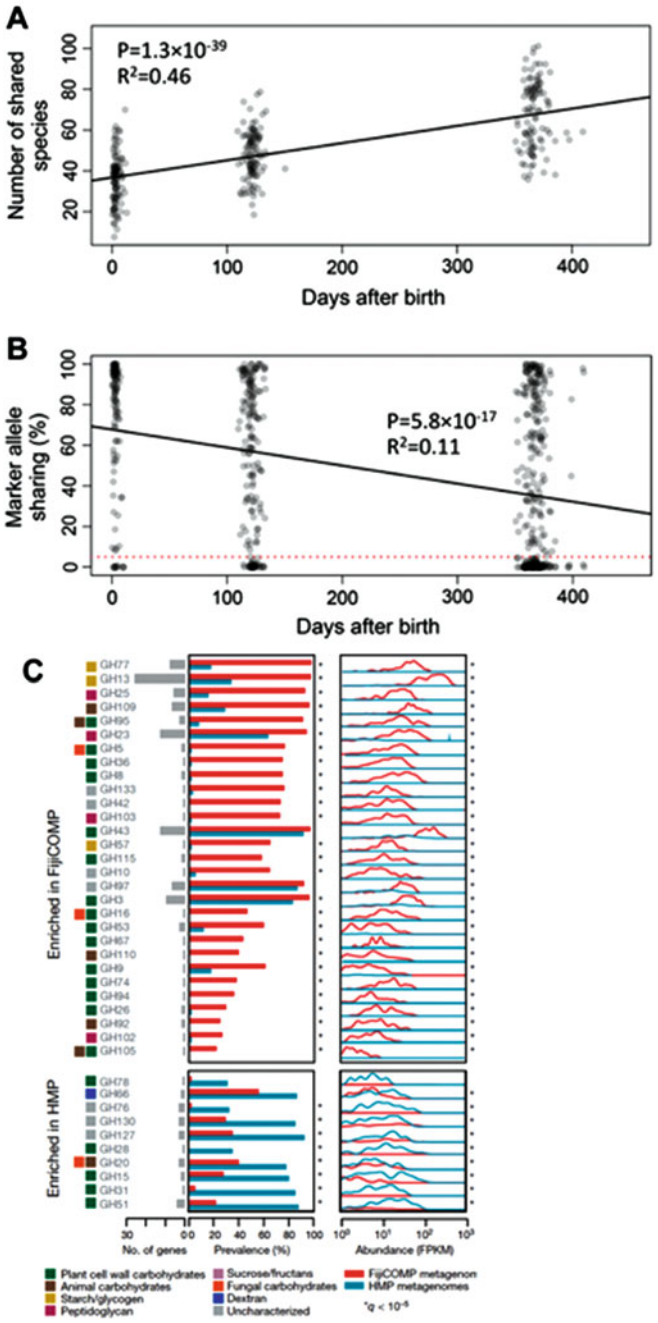


**Fig. 3** Examples of the automated resolution of strains in metagenomic data. (**a**, **b**) Comparison of metagenomic data in mother-infant pairs using MIDAS indicated that while the number of shared species increases with time after birth (**a**), vertical transmission of strains is particularly important

combination with automated tools greatly facilitates the use of metagenomic data to perform epidemiological studies. This allows us to expand on the studies that were thus far limited to isolate sequence data and enables insights into strain transmission, retention, and tissue specificity within the human body in the absence of any culturing bias (Donati et al. 2016).

Outside of the human microbiome, we are currently limited to analyzing a handful of lineages that have an adequate representation in the databases, although the generation of novel genomes reconstructed from metagenomic data and single cell genomics is rapidly increasing the number and taxonomic coverage of available references. For two well-represented taxa, the marine *Pelagibacter* and *Prochlorococcus*, MIDAS has been used to determine differences between populations in different oceanic regions by evaluating gene content overlap (Nayfach et al. 2016). Conventional approaches failed to detect these phenomena (Sunagawa et al. 2015). Whether these patterns were due to dispersal limitation or due to environmental selection according to conditions that differ between oceanic regions and that correlated with distance could not be resolved.

Expanding population-resolved analyses beyond taxa currently well represented in genomic databases, Garcia and coauthors generated their own system-specific (*in casu*, a specific lake) database of 33 reference genomes using single cell genomics and did read recruitment using metagenomic data from a 5-year sample time series from the same lake. They revealed distinct patterns for several abundant lineages. Some lineages could be resolved into distinct genotypes with clearly distinguished ecological dynamics that likely represented separate populations (e.g., *Actinobacteria* acI lineages). Other lineages (e.g., *Alphaproteobacteria* LD12, the freshwater sister group to marine *Pelagibacter*) did not have sequence-discrete nor ecologically distinct within-group dynamics (Garcia et al. 2016); thus distinct populations could not be resolved, or all sampled cells belonged to a single population. The availability of more nonhuman microbiome reference sequences, in combination with the recently developed automated tools to deconvolute strain patterns and identify alleles and gene content differences associated with these strains, is promising.

**Fig. 3** (continued) early in life and decreases as time goes on, based on the % of shared alleles in core genome marker genes (**b**). (**c**) Resolution of strain-specific differences due to divergence in mobile element gene content showed that the type and abundance of specific glycoside hydrolase gene families diverged significantly between a cohort from Fiji (FijiCOMP) and North America (HMP). Prevalence indicates the % of fecal samples in the cohort that the protein family was identified in. Abundance, expressed in fragments per kilobase of protein coding sequence per million mapped reads (FPKM), presents the relative abundance spectrum across all samples in each cohort. Asterisks indicate significant differences in prevalence and abundance. Figure adapted from Nayfach et al. (2016) Fig. 3 and Brito et al. (2016) Fig. 1

### 3.3 Determining Physiology of Ecologically and Genetically Cohesive Populations

In contrast to tracking population dynamics of closely related genotypes, only limited exploration of their physiological similarities and differences in the environment has been performed. When cultured isolates are available, it has been shown that closely related strains can adopt widely divergent physiologies, e.g., based on light spectrum preferences (Moore and Chisholm 1999) or temperature (Yung et al. 2015). Similar to metagenomics, a culture-independent approach can be taken to determine physiology of strains directly in the environment. This could theoretically be done by a combination of in situ hybridization [e.g., targeting genes sufficiently divergent to enable strain-specific hybridization using fluorescent in situ hybridization (Barrero-Canosa et al. 2017)] with assays gathering insights on physiology such as Raman spectrometry (Huang et al. 2007) or nano-SIMS (Behrens et al. 2008) that determine the ability for specific substrate uptake and/or metabolism.

Thus far, however, inferences about physiological differences between closely related but ecologically/genetically distinct populations have been made primarily by determining differences in transcript or protein abundances using metatranscriptomic and metaproteomic approaches. While translating gene expression to process rates remains challenging, recent studies integrating in situ expression and process measurements indicate the possibility to use gene expression data for process rate predictions (Wilson et al. 2017). Resolving expression patterns between closely related populations is particularly insightful when they are sympatric as these data can provide clues to the genetic differences that underlie ecological differences between these populations. Examples include the use of strain-resolved proteomics to show strain-level differences in biological phosphorus removal bioreactor communities (Wilmes et al. 2008), to identify pathways underpinning r- vs K-strategy ecotypes in biofilm development (Denef et al. 2010a), and to show physiological differences in chemotaxis and motility between closely related strains with distinct successional dynamics during preterm infant gut colonization (Brooks et al. 2015). All of these studies relied on a relatively labor-intensive manual effort to resolve strain-specific protein abundance levels and typically are focused on a single "species"-level group. More recently, automated strain-resolved metagenomic methods have also been used at a metatranscriptomic level [PanPhlAn; (Scholz et al. 2016)]. The method has been focused mostly on confirming activity of organisms in situ at strain-level resolution, for example, to show that strains vertically transmitted from mother to child were active in both the mother and child's gut environments (Asnicar et al. 2017).

Innovative tools have also been developed to gain insights into in situ growth rates of natural populations. When recruiting sequencing reads to assembled contigs from metagenomic data, it becomes apparent that replicating bacterial populations generate distinct coverage trends during bidirectional genome replication. Coverage is higher at the origin of replication and decreases toward the terminus. iRep is a tool that exploits this pattern to estimate an index of replication, which can be interpreted

as the fraction of the population that is actively making one genome copy at the time of sampling (Brown et al. 2016). The iRep estimate is a population-average value, and the existence of multiple replication forks during genome replication can bias this index (i.e., values >2 can be achieved). Olm and coauthors used iRep to track growth rates of strains across different body sites of preterm infants. First, they determined that identical strains could be found on multiple body parts. However, using iRep, they found that the replication rates of each strain differed depending on body site (Olm et al. 2017).

## 3.4 *Tracking Evolutionary Processes Within Ecologically and Genetically Cohesive Populations*

As stated at the start of the chapter, the analysis of metagenomic data allows us to resolve the genetic structure of natural populations. We discuss here findings related to the role of homologous recombination relative to mutation, variability in the flexible genome, and using metagenomic data to study natural selection.

Metagenomic analyses of the genetic structure of natural populations have led to new insights regarding the importance of homologous recombination within and between natural populations. Manual inspection of some of the first genomic datasets reconstructed from metagenomic data from an acid mine drainage system revealed the coexistence of multiple *Ferroplasma* populations that were inferred to be mosaic genomes originating from homologous recombination between at least three parent populations (Tyson et al. 2004). These findings were confirmed when comparing environmental metagenomic data to the genome of an isolate of the same species (Allen et al. 2007). A more quantitative approach was applied by Eppley and coauthors who found that the recombination rate within a *Ferroplasma* population was higher than the recombination rate between *Ferroplasma* populations. This suggested the presence of a species boundary based on genetic distance and within-species genetic cohesion mediated by homologous recombination (Eppley et al. 2007). Nonetheless, recombination still occurred between the two *Ferroplasma* populations, at rates proportional to varying sequence similarity across the genome. The continuation of homologous recombination in more conserved regions of the genome, while more divergent regions being already genetically more isolated, is in line with the model of temporally fragmented speciation proposed by Retchless and Lawrence (2007).

While all the studies mentioned in the previous paragraph focused on the same populations in acid mine drainage systems, they inspired new research on the importance of recombination in other systems and the development of automated methods to estimate recombination rates while controlling for sequencing errors (Johnson and Slatkin 2009). Subsequent studies found recombination to be common in marine populations, though at rates roughly four times lower than those observed in the acid mine drainage system archaeal populations (Konstantinidis and DeLong

2008). In thermophilic cyanobacteria, recombination rates have been shown to be similar to mutation rates observed through comparing metagenomic data with isolate genome sequence data (Rosen et al. 2015). In contrast, very low recombination rates relative to mutation rates were observed when comparing single cell genomes of LD12, the freshwater sister lineage of the abundant marine group *Pelagibacter* (Zaremba-Niedzwiedzka et al. 2013). These results indicate that recombination rates can be highly population-specific, and no generalization regarding the importance of recombination relative to mutation should be made. At the same time, it has to be noted that these rate comparisons generally do not control explicitly for differences in genetic distance between the sequences (and corresponding strains) considered.

Metagenomic data has also been used to study recombination within viral populations. A particular focus has been put on the CRISPR locus, which primarily functions as an adaptive defense system against viruses and is composed of an array of repeats interspersed with unique DNA segments called spacers. These CRISPR spacers most likely originate from the DNA of viruses infecting the microbial host that carries the CRISPR array in its DNA. Sequence reads that contained a sequence identical to a spacer sequence but no CRISPR repeats were identified as belonging to the targeted viruses and subsequently used to reconstruct viral genomic datasets. These reconstructions indicated the ability of some viruses to escape the microbial host's CRISPR viral defense system by homologous recombination. Erosion of linkage between viral genome variant positions at sequence lengths similar to the size of the CRISPR spacers leads to evasion of the CRISPR defense system by the viruses (Andersson and Banfield 2008). Similarly, by introducing multiple phage genotypes in a phage-bacterial coevolution experiment, recombination was shown to be an important mechanism to overcome CRISPR-based immunity (Paez-Espino et al. 2015).

Gene content differences between and within ecologically cohesive populations are observed commonly in studies using isolates. The analysis of metagenomic data has made it abundantly clear that genomic heterogeneity at the level of gene content is a hallmark of natural populations due to rapid gene gain and loss (Wilmes et al. 2009). The benefit of metagenomic data is that is has allowed for a quantitative assessment of the differential abundance of particular genomic islands between divergent environments (Coleman and Chisholm 2010) and over time (Bendall et al. 2016). The evolutionary origin of these gene content differences has been hypothesized to lie in a variety of ecological interactions (Cordero and Polz 2014) including viral predation (Rodriguez-Valera et al. 2009). From the enrichment of nutrient uptake genes under nutrient limitation (Coleman and Chisholm 2010), to the extensive gene flux in the mobile gene pool within and across species boundaries (Boucher et al. 2011), gene content differences are commonly observed to differentiate populations across space or time, despite overall cohesion of the rest of the genome.

Finally, efforts have been focused on identifying the effects of selection, which has been reviewed previously (Wilmes et al. 2009). Since that review, deep sampling of natural populations with metagenomic data generated from time series from a

relatively closed system (acid mine drainage) has been used to determine nucleotide fixation rates in the environment (Denef and Banfield 2012). The estimated rate was similar to findings in laboratory experimental evolution experiments (Barrick et al. 2009). Also, the loci affected by fixed non-synonymous mutations were biased toward regulatory genes in both the laboratory and environmental studies (Barrick et al. 2009), pointing to the importance of gene expression evolution in the early stages of evolutionary and ecological differentiation.

Despite challenges posed by dispersal in more open systems, a recent application of time-series metagenomics in a freshwater lake was able to show that both gene-specific and genome-wide selective sweeps occur in natural populations (Bendall et al. 2016) (Fig. 4). Other studies using isolates have indicated the possibility of gene-specific selective sweeps as well (Shapiro et al. 2012), and a previous metagenomic study has shown that orthologous regions differentiating coexisting organisms based on nucleotide substitutions did not show evidence of positive selection, contrary to predictions from the ecotype model (Simmons et al. 2008). Thus, population genomic studies using metagenomic data have added support for the importance of both gene-specific selective sweeps and genome-wide selective sweeps. The latter are in support of the ecotype model, i.e., that all diversity in an ecologically and evolutionary cohesive cluster of cells is regularly purged by selection of one specific adaptive genotype within the cluster, while the former indicates that recombination rates can be sufficiently high to undo the effects of selection. As argued by Shapiro and Polz (2015), there likely is no single model of speciation, but rather a spectrum determined by the contributions from gene flow and selection.

Time series metagenomic analyses have also suggested genome-wide selective sweeps in viral populations (Roux et al. 2014). Moreover, the dynamic interplay between viral and bacterial evolution has attracted the attention of researchers applying metagenomic tools, with a particular focus on dynamic changes occurring as a result of selection at CRISPR viral defense system loci. These analyses have given us insight into individual cell lineages' exposure history to viruses and have shown that CRISPR loci can be a population genome's most highly diverse loci (Tyson and Banfield 2008). Time series analyses of CRISPR sites have been used to (1) determine the retention of spacers and changes occurring in both CRISPR spacers and targeted viral genome loci (Sun et al. 2016), (2) model the evolutionary benefits of conservation of trailer-end (i.e., older) CRISPR spacers (Weinberger et al. 2012), and (3) identify molecular mechanisms such as incomplete immunity based on a single CRISPR spacer that may explain coevolutionary dynamics that deviate from those predicted by basic CRISPR immunity phage-bacteria population models (Levin et al. 2013). Similar datasets could be used to test recently proposed models of the dynamic coevolution between hosts and viruses based on CRISPR immunity (Childs et al. 2012).

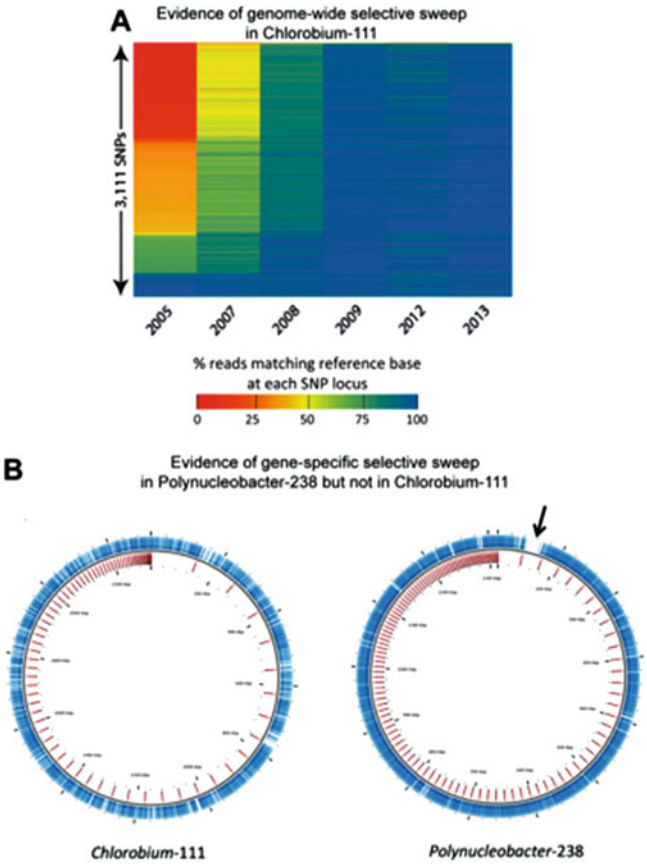


**Fig. 4** Identifying selection events using metagenomic-based population genomic analyses. (**a**) Read recruitment of metagenomic data generated from samples collected from the same lake over an 8-year period to a MAG of *Chlorobium*-111 indicated gradual purging of diversity at all polymorphic sites, i.e., a genome-wide selective sweep. The "reference base" is the base most commonly observed in the final sample in 2013. Data from samples from each year were combined for read recruitment. (**b**) Comparison of distribution SNPs (blue bars) detected in the metagenomic data across the MAGs of *Chlorobium*-111 and *Polynucleobacter*-238. Contigs breaks are indicated by red lines. The *Polynucleobacter* genome shows a 21 kbp region with no SNPs (black arrow), which the authors interpreted as evidence of a gene-specific selective sweep preceding the first sample time point. Adapted from Bendall et al. (2016) Fig. 4 and Supplementary Figure S5

## 4 Outlook

Despite tremendous insights into natural population genomic heterogeneity gained from metagenomic approaches, some of the key limitations of metagenomic data have kept metagenomics from replacing isolate or single cell genomic approaches to perform population genomic analyses. This is particularly true for the estimation of

key population genetic parameters. While future advances in read length and base calling accuracies may facilitate the use of metagenomic data for population genomic analyses sensu stricto (Koren and Phillippy 2015), recently developed tools discussed in this chapter are allowing us to mine current metagenomic data to identify and track strains across space and time (Asnicar et al. 2017; Nayfach et al. 2016; Quince et al. 2017). As discussed above, such approaches are most powerful in the context of extensive reference genomic databases, making them currently most useful in human microbiome research. Yet, the ability to readily obtain (partial) genomic sequences from 100 s to 1000 s of single cells per sample (Kashtan et al. 2014, 2017) or directly from metagenomic data (Anantharaman et al. 2016; Delmont et al. 2017) is opening avenues to apply these tools to all microbial systems. We will likely also see further integration of population genomic analyses with metatranscriptomic or metaproteomic data or even high-throughput measurements of phenotypic features (Props et al. 2016) to gain insights into both the role of within- and between-population genomic heterogeneity and phenotypic plasticity. Improving our ability to see changes in population genetic structure of microbial populations across space and time will improve our understanding of both the evolutionary and the ecological processes that shape microbial populations (Cohan 2016; Dudaniec and Tesson 2016; Shapiro and Polz 2015). These insights are paramount in our efforts to understand how microbial populations and the communities they are part of change in composition and functioning in light of change, particularly disturbances caused by human activities.

**Acknowledgments** I thank Ruben Props (Ghent University) and Prof. Martin Polz (MIT) for constructive comments to help improve this chapter.

## References

Acinas SG, Klepac-Ceraj V, Hunt DE, Pharino C, Ceraj I, Distel DL, Polz MF. Fine-scale phylogenetic architecture of a complex bacterial community. Nature. 2004;430:551–4.

Albertsen M, Hugenholtz P, Skarshewski A, Nielsen KL, Tyson GW, Nielsen PH. Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. Nat Biotechnol. 2013;31:533–8.

Allen EE, Tyson GW, Whitaker RJ, Detter JC, Richardson PM, Banfield JF. Genome dynamics in a natural archaeal population. Proc Natl Acad Sci U S A. 2007;104:1883–8.

Alneberg J, Bjarnason BS, de Bruijn I, Schirmer M, Quick J, Ijaz UZ, Lahti L, Loman NJ, Andersson AF, Quince C. Binning metagenomic contigs by coverage and composition. Nat Methods. 2014;11:1144–6.

Anantharaman K, Brown CT, Hug LA, Sharon I, Castelle CJ, Probst AJ, Thomas BC, Singh A, Wilkins MJ, Karaoz U. Thousands of microbial genomes shed light on interconnected biogeochemical processes in an aquifer system. Nature Commun. 2016;7:13219.

Andersson AF, Banfield JF. Virus population dynamics and acquired virus resistance in natural microbial communities. Science. 2008;320:1047–50.

Asnicar F, Manara S, Zolfo M, Truong DT, Scholz M, Armanini F, Ferretti P, Gorfer V, Pedrotti A, Tett A, Segata N. Studying vertical microbiome transmission from mothers to infants by strain-

level metagenomic profiling mSystems. 2017;2(1). pii: e00164-16. doi: https://doi.org/10.1128/mSystems.00164-16.
Baker BJ, Tyson GW, Webb RI, Flanagan J, Hugenholtz P, Allen EE, Banfield JF. Lineages of acidophilic archaea revealed by community genomic analysis. Science. 2006;314:1933–5.
Baker BJ, Comolli LR, Dick GJ, Hauser LJ, Hyatt D, Dill BD, Land ML, Verberkmoes NC, Hettich RL, Banfield JF. Enigmatic, ultrasmall, uncultivated Archaea. Proc Natl Acad Sci U S A. 2010;107:8806–11.
Barrero-Canosa J, Moraru C, Zeugner L, Fuchs BM, Amann R. Direct-geneFISH: a simplified protocol for the simultaneous detection and quantification of genes and rRNA in microorganisms. Environ Microbiol. 2017;19:70–82.
Barrick JE, Yu DS, Yoon SH, Jeong H, Oh TK, Schneider D, Lenski RE, Kim JF. Genome evolution and adaptation in a long-term experiment with Escherichia coli. Nature. 2009;461:1243–7.
Behrens S, Lösekann T, Pett-Ridge J, Weber PK, Ng WO, Stevenson BS, Hutcheon ID, Relman DA, Spormann AM. Linking microbial phylogeny to metabolic activity at the single-cell level by using enhanced element labeling-catalyzed reporter deposition fluorescence in situ hybridization (EL-FISH) and NanoSIMS. Appl Environ Microbiol. 2008;74:3143–50.
Bendall ML, Stevens SL, Chan LK, Malfatti S, Schwientek P, Tremblay J, Schackwitz W, Martin J, Pati A, Bushnell B, Froula J, Kang D, Tringe SG, Bertilsson S, Moran MA, Shade A, Newton RJ, McMahon KD, Malmstrom RR. Genome-wide selective sweeps and gene-specific sweeps in natural bacterial populations. ISME J. 2016;10:1589–601.
Berry MA, White JD, Davis TW, Jain S, Johengen TH, Dick GJ, Sarnelle O, Denef VJ. Are oligotypes meaningful ecological and phylogenetic units? A case study of Microcystis in freshwater lakes. Front Microbiol. 2017;8:365.
Bhaya D, Grossman AR, Steunou AS, Khuri N, Cohan FM, Hamamura N, Melendrez MC, Bateson MM, Ward DM, Heidelberg JF. Population level functional diversity in a microbial community revealed by comparative genomic and metagenomic analyses. ISME J. 2007;1:703–13.
Blainey PC. The future is now: single-cell genomics of bacteria and archaea. FEMS Microbiol Rev. 2013;37:407–27.
Boucher Y, Cordero OX, Takemura A, Hunt DE, Schliep K, Bapteste E, Lopez P, Tarr CL, Polz MF. Local mobile gene pools rapidly cross species boundaries to create endemicity within global Vibrio cholerae populations. MBio 2011;2(2). pii: e00335-10. doi: https://doi.org/10.1128/mBio.00335-10
Brito IL, Yilmaz S, Huang K, Xu L, Jupiter SD, Jenkins AP, Naisilisili W, Tamminen M, Smillie CS, Wortman JR, Birren BW, Xavier RJ, Blainey PC, Singh AK, Gevers D, Alm EJ. Mobile genes in the human microbiome are structured from global to individual scales. Nature. 2016;535:435–9.
Broeksema B, Calusinska M, McGee F, Winter K, Bongiovanni F, Goux X, Wilmes P, Delfosse P, Ghoniem M. ICoVeR—an interactive visualization tool for verification and refinement of metagenomic bins. BMC Bioinformatics. 2017;18:233.
Brooks B, Mueller RS, Young JC, Morowitz MJ, Hettich RL, Banfield JF. Strain-resolved microbial community proteomics reveals simultaneous aerobic and anaerobic function during gastrointestinal tract colonization of a preterm infant. Front Microbiol. 2015;6:654.
Brown CT, Olm MR, Thomas BC, Banfield JF. Measurement of bacterial replication rates in microbial communities. Nat Biotechnol. 2016;34:1256–63.
Caro-Quintero A, Konstantinidis KT. Bacterial species may exist, metagenomics reveal. Environ Microbiol. 2012;14:347–55.
Childs LM, Held NL, Young MJ, Whitaker RJ, Weitz JS. Multiscale model of CRISPR-induced coevolutionary dynamics: diversification at the interface of Lamarck and Darwin. Evolution. 2012;66:2015–29.
Cleary B, Brito IL, Huang K, Gevers D, Shea T, Young S, Alm EJ. Detection of low-abundance bacterial strains in metagenomic datasets by eigengenome partitioning. Nat Biotechnol. 2015;33:1053–60.

Cohan FM. Bacterial speciation: genetic sweeps in bacterial species. Curr Biol. 2016;26:R112–5.
Coleman ML, Chisholm SW. Ecosystem-specific selection pressures revealed through comparative population genomics. Proc Natl Acad Sci U S A. 2010;107:18634–9.
Coleman ML, Sullivan MB, Martiny AC, Steglich C, Barry K, Delong EF, Chisholm SW. Genomic islands and the ecology and evolution of Prochlorococcus. Science. 2006;311:1768–70.
Cordero OX, Polz MF. Explaining microbial genomic diversity in light of evolutionary ecology. Nat Rev Microbiol. 2014;12:263–73.
Cordero OX, Ventouras LA, DeLong EF, Polz MF. Public good dynamics drive evolution of iron acquisition strategies in natural bacterioplankton populations. Proc Natl Acad Sci U S A. 2012;109:20059–64.
Delmont TO, Eren AM, Maccario L, Prestat E, Esen ÖC, Pelletier E, Le Paslier D, Simonet P, Vogel TM. Reconstructing rare soil microbial genomes using in situ enrichments and metagenomics. Front Microbiol. 2015;6:358.
Delmont TO, Quince C, Shaiber A, Esen OC, Lee ST, Lucker S, Eren AM. Nitrogen-fixing populations of Planctomycetes and Proteobacteria are abundant in the surface ocean. bioRxiv. 2017:129791.
DeLong EF. Microbial population genomics and ecology: the road ahead. Environ Microbiol. 2004;6:875–8.
DeLong EF. Microbial community genomics in the ocean. Nat Rev Microbiol. 2005;3:459–69.
DeLong EF. Microbial evolution in the wild. Science. 2012;336:422–4.
Denef VJ, Banfield JF. In situ evolutionary rate measurements show ecological success of recently emerged bacterial hybrids. Science. 2012;336:462–6.
Denef VJ, Kalnejais LH, Mueller RS, Wilmes P, Baker BJ, Thomas BC, VerBerkmoes NC, Hettich RL, Banfield JF. Proteogenomic basis for ecological divergence of closely related bacteria in natural acidophilic microbial communities. Proc Natl Acad Sci U S A. 2010a;107:2383–90.
Denef VJ, Mueller RS, Banfield JF. AMD biofilms: using model communities to study microbial evolution and ecological complexity in nature. ISME J. 2010b;4:599–610.
Deng L, Ignacio-Espinoza JC, Gregory AC, Poulos BT, Weitz JS, Hugenholtz P, Sullivan MB. Viral tagging reveals discrete populations in Synechococcus viral genome sequence space. Nature. 2014;513:242–5.
Dick GJ, Andersson AF, Baker BJ, Simmons SL, Thomas BC, Yelton AP, Banfield JF. - Community-wide analysis of microbial genome sequence signatures. Genome Biol. 2009;10: R85.
Donati C, Zolfo M, Albanese D, Tin Truong D, Asnicar F, Iebba V, Cavalieri D, Jousson O, De Filippo C, Huttenhower C, Segata N. Uncovering oral Neisseria tropism and persistence using metagenomic sequencing. Nat Microbiol. 2016;1:16070.
Dudaniec RY, Tesson SV. Applying landscape genetics to the microbial world. Mol Ecol. 2016;25:3266–75.
Eckburg PB, Bik EM, Bernstein CN, Purdom E, Dethlefsen L, Sargent M, Gill SR, Nelson KE, Relman DA. Diversity of the human intestinal microbial flora. Science. 2005;308:1635–8.
Eppley JM, Tyson GW, Getz WM, Banfield JF. Genetic exchange across a species boundary in the archaeal genus ferroplasma. Genetics. 2007;177:407–16.
Eren AM, Maignien L, Sul WJ, Murphy LG, Grim SL, Morrison HG, Sogin ML. Oligotyping: differentiating between closely related microbial taxa using 16S rRNA gene data. Methods Ecol Evol. 2013;4
Eren AM, Esen C, Quince C, Vineis JH, Morrison HG, Sogin ML, Delmont TO. Anvio: an advanced analysis and visualization platform for omics data. PeerJ. 2015;3:e1319.
Fraser C, Alm EJ, Polz MF, Spratt BG, Hanage WP. The bacterial species challenge: making sense of genetic and ecological diversity. Science. 2009;323:741–6.
Frias-Lopez J, Shi Y, Tyson GW, Coleman ML, Schuster SC, Chisholm SW, DeLong EF. Microbial community gene expression in ocean surface waters. Proc Natl Acad Sci U S A. 2008;105:3805–10.
Fuhrman JA, Campbell L. Marine ecology: microbial microdiversity. Nature. 1998;393:410–1.

Garcia SL, Stevens SL, Crary B, Martinez-Garcia M, Stepanauskas R, Woyke T, Tringe SG, Andersson S, Bertilsson S, Malmstrom RR. Contrasting patterns of genome-level diversity across distinct co-occurring bacterial populations. bioRxiv. 2016. https://doi.org/10.1101/080168.

Giovannoni SJ, Britschgi TB, Moyer CL, Field KG. Genetic diversity in Sargasso Sea bacterioplankton. Nature. 1990;345:60.

Gogarten JP, Townsend JP. Horizontal gene transfer, genome innovation and evolution. Nat Rev Microbiol. 2005;3:679–87.

Hahn MW, Jezberová J, Koll U, Saueressig-Beck T, Schmidt J. Complete ecological isolation and cryptic diversity in Polynucleobacter bacteria not resolved by 16S rRNA gene sequences. ISME J. 2016;10:1642–55.

Hanage WP, Fraser C, Spratt BG. Sequences, sequence clusters and bacterial species. Phil Trans R Soc B. 2006;361:1917–27.

Handelsman J. Metagenomics: application of genomics to uncultured microorganisms. Microbiol Mol Biol Rev. 2004;68:669–85.

Hatzenpichler R, Connon SA, Goudeau D, Malmstrom RR, Woyke T, Orphan VJ. Visualizing in situ translational activity for identifying and sorting slow-growing archaeal-bacterial consortia. Proc Natl Acad Sci U S A. 2016;113:E4069–78.

Haubold B, Pfaffelhuber P, Lynch M. mlRho—a program for estimating the population mutation and recombination rates from shotgun-sequenced diploid genomes. Mol Ecol. 2010;19(Suppl 1):277–84.

Hehemann JH, Arevalo P, Datta MS, Yu X, Corzett CH, Henschel A, Preheim SP, Timberlake S, Alm EJ, Polz MF. Adaptive radiation by waves of gene transfer leads to fine-scale resource partitioning in marine microbes. Nat Commun. 2016;7:12860.

Huang WE, Stoecker K, Griffiths R, Newbold L, Daims H, Whiteley AS, Wagner M. Raman-FISH: combining stable-isotope Raman spectroscopy and fluorescence in situ hybridization for the single cell analysis of identity and function. Environ Microbiol. 2007;9:1878–89.

Hunt DE, David LA, Gevers D, Preheim SP, Alm EJ, Polz MF. Resource partitioning and sympatric differentiation among closely related bacterioplankton. Science. 2008;320:1081–5.

Imelfort M, Parks D, Woodcroft BJ, Dennis P, Hugenholtz P, Tyson GW. GroopM: an automated tool for the recovery of population genomes from related metagenomes. PeerJ. 2014;2:e603.

Johnson PL, Slatkin M. Inference of population genetic parameters in metagenomics: a clean look at messy data. Genome Res. 2006;16:1320–7.

Johnson PL, Slatkin M. Accounting for bias from sequencing error in population genetic estimates. Mol Biol Evol. 2008;25:199–206.

Johnson PL, Slatkin M. Inference of microbial recombination rates from metagenomic data. PLoS Genet. 2009;5:e1000674.

Kashtan N, Roggensack SE, Rodrigue S, Thompson JW, Biller SJ, Coe A, Ding H, Marttinen P, Malmstrom RR, Stocker R, Follows MJ, Stepanauskas R, Chisholm SW. Single-cell genomics reveals hundreds of coexisting subpopulations in wild Prochlorococcus. Science. 2014;344:416–20.

Kashtan N, Roggensack SE, Berta-Thompson JW, Grinberg M, Stepanauskas R, Chisholm SW. Fundamental differences in diversity and genomic population structure between Atlantic and Pacific Prochlorococcus. ISME J. 2017;11(9):1997–2011.

Konstantinidis KT, DeLong EF. Genomic patterns of recombination, clonal divergence and environment in marine microbial populations. ISME J. 2008;2:1052–65.

Koren S, Phillippy AM. One chromosome, one contig: complete microbial genomes from long-read sequencing and assembly. Curr Opin Microbiol. 2015;23:110–20.

Krause DJ, Whitaker RJ. Inferring speciation processes from patterns of natural variation in microbial genomes. Syst Biol. 2015;64:926–35.

Kuo C-H, Ochman H. The fate of new bacterial genes. FEMS Microbiol Rev. 2009;33:38–43.

Laczny CC, Sternal T, Plugaru V, Gawron P, Atashpendar A, Margossian HH, Coronado S, Lv d M, Vlassis N, Wilmes P. VizBin—an application for reference-independent visualization and human-augmented binning of metagenomic data. Microbiome. 2015;3(1):1.

Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9:357–9.

Larkin AA, Martiny AC. Microdiversity shapes the traits, niche space, and biogeography of microbial taxa. Environ Microbiol Rep. 2017;9:55–70.

Levin BR, Moineau S, Bushman M, Barrangou R. The population and evolutionary dynamics of phage and bacteria with CRISPR-mediated immunity. PLoS Genet. 2013;9:e1003312.

Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics. 2010;26:589–95.

Luo C, Knight R, Siljander H, Knip M, Xavier RJ, Gevers D. ConStrains identifies microbial strains in metagenomic datasets. Nat Biotechnol. 2015;33:1045–52.

Maiden MC, Bygraves JA, Feil E, Morelli G, Russell JE, Urwin R, Zhang Q, Zhou J, Zurth K, Caugant DA, Feavers IM, Achtman M, Spratt BG. Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. Proc Natl Acad Sci U S A. 1998;95:3140–5.

Malmstrom RR, Rodrigue S, Huang KH, Kelly L, Kern SE, Thompson A, Roggensack S, Berube PM, Henn MR, Chisholm SW. Ecology of uncultured Prochlorococcus clades revealed through single-cell genomics and biogeographic analysis. ISME J. 2013;7:184–98.

Marbouty M, Cournac A, Flot JF, Marie-Nelly H, Mozziconacci J, Koszul R. Metagenomic chromosome conformation capture (meta3C) unveils the diversity of chromosome organization in microorganisms. Elife. 2014;3:e03318.

Moore LR, Chisholm SW. Photophysiology of the marine cyanobacterium Prochlorococcus: Ecotypic differences among cultured isolates. Limnol Oceanogr. 1999;44:628–38.

Morowitz MJ, Denef VJ, Costello EK, Thomas BC, Poroyko V, Relman DA, Banfield JF. Strain-resolved community genomic analysis of gut microbial colonization in a premature infant. Proc Natl Acad Sci U S A. 2011;108:1128–33.

Nayfach S, Rodriguez-Mueller B, Garud N, Pollard KS. An integrated metagenomics pipeline for strain profiling reveals novel patterns of bacterial transmission and biogeography. Genome Res. 2016;26:1612–25.

Nielsen R, Paul JS, Albrechtsen A, Song YS. Genotype and SNP calling from next-generation sequencing data. Nat Rev Genet. 2011;12:443–51.

Nijkamp JF, Pop M, Reinders MJ, de Ridder D. Exploring variation-aware contig graphs for (comparative) metagenomics using MaryGold. Bioinformatics. 2013;29:2826–34.

Oh S, Caro-Quintero A, Tsementzi D, DeLeon-Rodriguez N, Luo C, Poretsky R, Konstantinidis KT. Metagenomic insights into the evolution, function, and complexity of the planktonic microbial community of Lake Lanier, a temperate freshwater ecosystem. Appl Environ Microbiol. 2011;77:6000–11.

Olm MR, Brown CT, Brooks B, Firek B, Baker R, Burstein D, Soenjoyo K, Thomas BC, Morowitz M, Banfield JF. Identical bacterial populations colonize premature infant gut, skin, and oral microbiomes and exhibit different in situ growth rates. Genome Res. 2017;27:601–12.

Paez-Espino D, Sharon I, Morovic W, Stahl B, Thomas BC, Barrangou R, Banfield JF. CRISPR immunity drives rapid phage genome evolution in Streptococcus thermophilus. MBio. 2015;6.

Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res. 2015;25:1043–55.

Parks DH, Rinke C, Chuvochina M, Chaumeil PA, Woodcroft BJ, Evans PN, Hugenholtz P, Tyson GW. Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. Nat Microbiol. 2017. https://doi.org/10.1038/s41564-017-0012-7.

Pernthaler A, Dekas AE, Brown CT, Goffredi SK, Embaye T, Orphan VJ. Diverse syntrophic partnerships from deep-sea methane vents revealed by direct cell capture and metagenomics. Proc Natl Acad Sci U S A. 2008;105:7052–7.

Props R, Monsieurs P, Mysara M, Clement L, Boon N. Measuring the biodiversity of microbial communities by flow cytometry. Meth Ecol Evol. 2016;7:1376–85.
Quince C, Delmont TO, Raguideau S, Alneberg J, Darling AE, Collins G, Eren AM. DESMAN: a new tool for de novo extraction of strains from metagenomes. Genome Biol. 2017;18:181.
Retchless AC, Lawrence JG. Temporal fragmentation of speciation in bacteria. Science. 2007;317:1093–6.
Rocap G, Larimer FW, Lamerdin J, Malfatti S, Chain P, Ahlgren NA, Arellano A, Coleman M, Hauser L, Hess WR, Johnson ZI, Land M, Lindell D, Post AF, Regala W, Shah M, Shaw SL, Steglich C, Sullivan MB, Ting CS, Tolonen A, Webb EA, Zinser ER, Chisholm SW. Genome divergence in two Prochlorococcus ecotypes reflects oceanic niche differentiation. Nature. 2003;424:1042–7.
Rodriguez-Brito B, Li L, Wegley L, Furlan M, Angly F, Breitbart M, Buchanan J, Desnues C, Dinsdale E, Edwards R, Felts B, Haynes M, Liu H, Lipson D, Mahaffy J, Martin-Cuadrado AB, Mira A, Nulton J, Pasić L, Rayhawk S, Rodriguez-Mueller J, Rodriguez-Valera F, Salamon P, Srinagesh S, Thingstad TF, Tran T, Thurber RV, Willner D, Youle M, Rohwer F. Viral and microbial community dynamics in four aquatic environments. ISME J. 2010;4:739–51.
Rodriguez-Valera F, Martin-Cuadrado AB, Rodriguez-Brito B, Pasić L, Thingstad TF, Rohwer F, Mira A. Explaining microbial population genomics through phage predation. Nat Rev Microbiol. 2009;7:828–36.
Rodriguez-Valera F, Martin-Cuadrado AB, López-Pérez M. Flexible genomic islands as drivers of genome evolution. Curr Opin Microbiol. 2016;31:154–60.
Rosen MJ, Davison M, Bhaya D, Fisher DS. Fine-scale diversity and extensive recombination in a quasisexual bacterial population occupying a broad niche. Science. 2015;348:1019–23.
Roux S, Hawley AK, Beltran MT, Scofield M, Schwientek P, Stepanauskas R, Woyke T, Hallam SJ, Sullivan MB. Ecology and evolution of viruses infecting uncultivated SUP05 bacteria as revealed by single-cell-and meta-genomics. Elife. 2014;3:e03125.
Schirmer M, D'Amore R, Ijaz UZ, Hall N, Quince C. Illumina error profiles: resolving fine-scale variation in metagenomic sequencing data. BMC Bioinformatics. 2016;17:125.
Scholz M, Ward DV, Pasolli E, Tolio T, Zolfo M, Asnicar F, Truong DT, Tett A, Morrow AL, Segata N. Strain-level microbial epidemiology and population genomics from shotgun metagenomics. Nat Methods. 2016;13:435–8.
Shapiro BJ. What microbial population genomics has taught us about speciation. Popul Genom. 2017. https://doi.org/10.1007/13836_2018_10.
Shapiro BJ, Polz MF. Ordering microbial diversity into ecologically and genetically cohesive units. Trends Microbiol. 2014;22:235–47.
Shapiro BJ, Polz MF. Microbial speciation. Cold Spring Harb Perspect Biol. 2015;7(10):a018143. https://doi.org/10.1101/cshperspect.a0181.
Shapiro BJ, Friedman J, Cordero OX, Preheim SP, Timberlake SC, Szabó G, Polz MF, Alm EJ. Population genomics of early events in the ecological differentiation of bacteria. Science. 2012;336:48–51.
Sharon I, Morowitz MJ, Thomas BC, Costello EK, Relman DA, Banfield JF. Time series community genomics analysis reveals rapid shifts in bacterial species, strains, and phage during infant gut colonization. Genome Res. 2013;23:111–20.
Sharon I, Kertesz M, Hug LA, Pushkarev D, Blauwkamp TA, Castelle CJ, Amirebrahimi M, Thomas BC, Burstein D, Tringe SG, Williams KH, Banfield JF. Accurate, multi-kb reads resolve complex populations and detect rare microorganisms. Genome Res. 2015;25:534–43.
Simmons SL, Dibartolo G, Denef VJ, Goltsman DS, Thelen MP, Banfield JF. Population genomic analysis of strain variation in Leptospirillum group II bacteria involved in acid mine drainage formation. PLoS Biol. 2008;6:e177.
Sun CL, Thomas BC, Barrangou R, Banfield JF. Metagenomic reconstructions of bacterial CRISPR loci constrain population histories. ISME J. 2016;10:858–70.

Sunagawa S, Coelho LP, Chaffron S, Kultima JR, Labadie K, Salazar G, Djahanschiri B, Zeller G, Mende DR, Alberti A. Structure and function of the global ocean microbiome. Science. 2015;348:1261359.

Thompson AW, Huang K, Saito MA, Chisholm SW. Transcriptome response of high-and low-light-adapted Prochlorococcus strains to changing iron availability. ISME J. 2011;5:1580–94.

Thrash JC, Temperton B, Swan BK, Landry ZC, Woyke T, DeLong EF, Stepanauskas R, Giovannoni SJ. Single-cell enabled comparative genomics of a deep ocean SAR11 bathytype. ISME J. 2014;8:1440–51.

Truong DT, Franzosa EA, Tickle TL, Scholz M, Weingart G, Pasolli E, Tett A, Huttenhower C, Segata N. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. Nat Methods. 2015;12:902–3.

Truong DT, Tett A, Pasolli E, Huttenhower C, Segata N. Microbial strain-level population structure and genetic diversity from metagenomes. Genome Res. 2017;27:626–38.

Tyson GW, Banfield JF. Rapidly evolving CRISPRs implicated in acquired resistance of microorganisms to viruses. Environ Microbiol. 2008;10:200–7.

Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, Richardson PM, Solovyev VV, Rubin EM, Rokhsar DS, Banfield JF. Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature. 2004;428:37–43.

Ward DV, Scholz M, Zolfo M, Taft DH, Schibler KR, Tett A, Segata N, Morrow AL. Metagenomic sequencing with strain-level resolution implicates uropathogenic E. coli in necrotizing enterocolitis and mortality in preterm infants. Cell Rep. 2016;14:2912–24.

Watterson GA. On the number of segregating sites in genetical models without recombination. Theor Popul Biol. 1975;7:256–76.

Weinberger AD, Sun CL, Pluciński MM, Denef VJ, Thomas BC, Horvath P, Barrangou R, Gilmore MS, Getz WM, Banfield JF. Persisting viral sequences shape microbial CRISPR-based immunity. PLoS Comput Biol. 2012;8:e1002475.

Welch RA, Burland V, Plunkett G, Redford P, Roesch P, Rasko D, Buckles EL, Liou SR, Boutin A, Hackett J, Stroud D, Mayhew GF, Rose DJ, Zhou S, Schwartz DC, Perna NT, Mobley HL, Donnenberg MS, Blattner FR. Extensive mosaic structure revealed by the complete genome sequence of uropathogenic Escherichia coli. Proc Natl Acad Sci U S A. 2002;99:17020–4.

Whitaker RJ, Banfield JF. Population genomics in natural microbial communities. Trends Ecol Evol. 2006;21:508–16.

Wilmes P, Andersson AF, Lefsrud MG, Wexler M, Shah M, Zhang B, Hettich RL, Bond PL, VerBerkmoes NC, Banfield JF. Community proteogenomics highlights microbial strain-variant protein expression within activated sludge performing enhanced biological phosphorus removal. ISME J. 2008;2:853–64.

Wilmes P, Simmons SL, Denef VJ, Banfield JF. The dynamic genetic repertoire of microbial communities. FEMS Microbiol Rev. 2009;33:109–32.

Wilson ST, Aylward FO, Ribalet F, Barone B, Casey JR, Connell PE, Eppley JM, Ferrón S, Fitzsimmons JN, Hayes CT, Romano AE, Turk-Kubo KA, Vislova A, Armbrust EV, Caron DA, Church MJ, Zehr JP, Karl DM, DeLong EF. Coordinated regulation of growth, activity and transcription in natural populations of the unicellular nitrogen-fixing cyanobacterium *Crocosphaera*. Nat Microbiol. 2017;2:17118.

Yung CM, Vereen MK, Herbert A, Davis KM, Yang J, Kantorowska A, Ward CS, Wernegreen JJ, Johnson ZI, Hunt DE. Thermally adaptive tradeoffs in closely related marine bacterial strains. Environ Microbiol. 2015;17:2421–9.

Zaremba-Niedzwiedzka K, Viklund J, Zhao W, Ast J, Sczyrba A, Woyke T, McMahon K, Bertilsson S, Stepanauskas R, Andersson SG. Single-cell genomics reveal low recombination frequencies in freshwater bacteria of the SAR11 clade. Genome Biol. 2013;14:R130.

Zojer M, Schuster LN, Schulz F, Pfundner A, Horn M, Rattei T. Variant profiling of evolving prokaryotic populations. PeerJ. 2017;5:e2997.

Zolfo M, Tett A, Jousson O, Donati C, Segata N. MetaMLST: multi-locus strain-level bacterial typing from metagenomic samples. Nucleic Acids Res. 2017;45:e7.