Indian J Microbiol (Jan–Mar 2017) 57(1):23–38
DOI 10.1007/s12088-016-0629-x

REVIEW ARTICLE

Survey of (Meta)genomic Approaches for Understanding Microbial Community Dynamics

Anukriti Sharma[1] · **Rup Lal**[1]

Received: 1 July 2016 / Accepted: 27 October 2016 / Published online: 11 November 2016

Abstract Advancement in the next generation sequencing technologies has led to evolution of the field of genomics and metagenomics in a slim duration with nominal cost at precipitous higher rate. While metagenomics and genomics can be separately used to reveal the culture-independent and culture-based microbial evolution, respectively, (meta)genomics together can be used to demonstrate results at population level revealing in-depth complex community interactions for specific ecotypes. The field of metagenomics which started with answering "who is out there?" based on 16S rRNA gene has evolved immensely with the precise organismal reconstruction at species/strain level from the deeply covered metagenome data outweighing the need to isolate bacteria of which 99% are de facto non-cultivable. In this review we have underlined the appeal of metagenomic-derived genomes in providing insights into the evolutionary patterns, growth dynamics, genome/gene-specific sweeps, and durability of environmental pressures. We have demonstrated the use of culture-based genomics and environmental shotgun metagenome data together to elucidate environment specific genome modulations via metagenomic recruitments in terms of gene loss/gain, accessory and core-genome extent. We further illustrated the benefit of (meta)genomics in the understanding of infectious diseases by deducing the relationship between human microbiota and clinical microbiology. This review summarizes the technological advances in the (meta)genomic strategies using the genome and metagenome datasets together to increase the resolution of microbial population studies.

Keywords Metagenomics · Genomics · De novo · Genome reconstruction · Recruitments

Introduction

With the advent of next generation sequencing (NGS) technologies, the field of (meta)genomics has revolutionized the landscape of microbiology leading to deluge of environmental and genome sequence data. With the availability of sequenced bacterial genomes, comparative genomics has emerged to be indispensable in elucidating evolutionary forces active across genera or species; however it remains incompetent to demonstrate results at the population level of in situ cohorts in an environment [1–11]. The problem can be resolved to some extent by using genomics and metagenomics data together delineating the pan-genome dynamics of a community distinctly elucidating environment specific lifestyles adopted by bacteria (Fig. 1) [12, 13].

Ever since its discovery 'metagenomics' has been largely used to decipher the overall taxonomic composition at an environment focusing more on 'meta' and less on 'genomics' part except for the organismal reconstruction at the species/strain or pan-genome level (multiple species of a specific genus) [12, 14]. While there are numerous studies based on using genomics and metagenomics independently, it has not been until recently that the potential of mining genome and metagenome datasets together was exploited to unveil complex environment–host interactions. Using (meta)genomics in sync, can provide a better understanding of habitat independent

✉ Rup Lal
ruplal@gmail.com

[1] Department of Zoology, University of Delhi, Delhi 110007, India

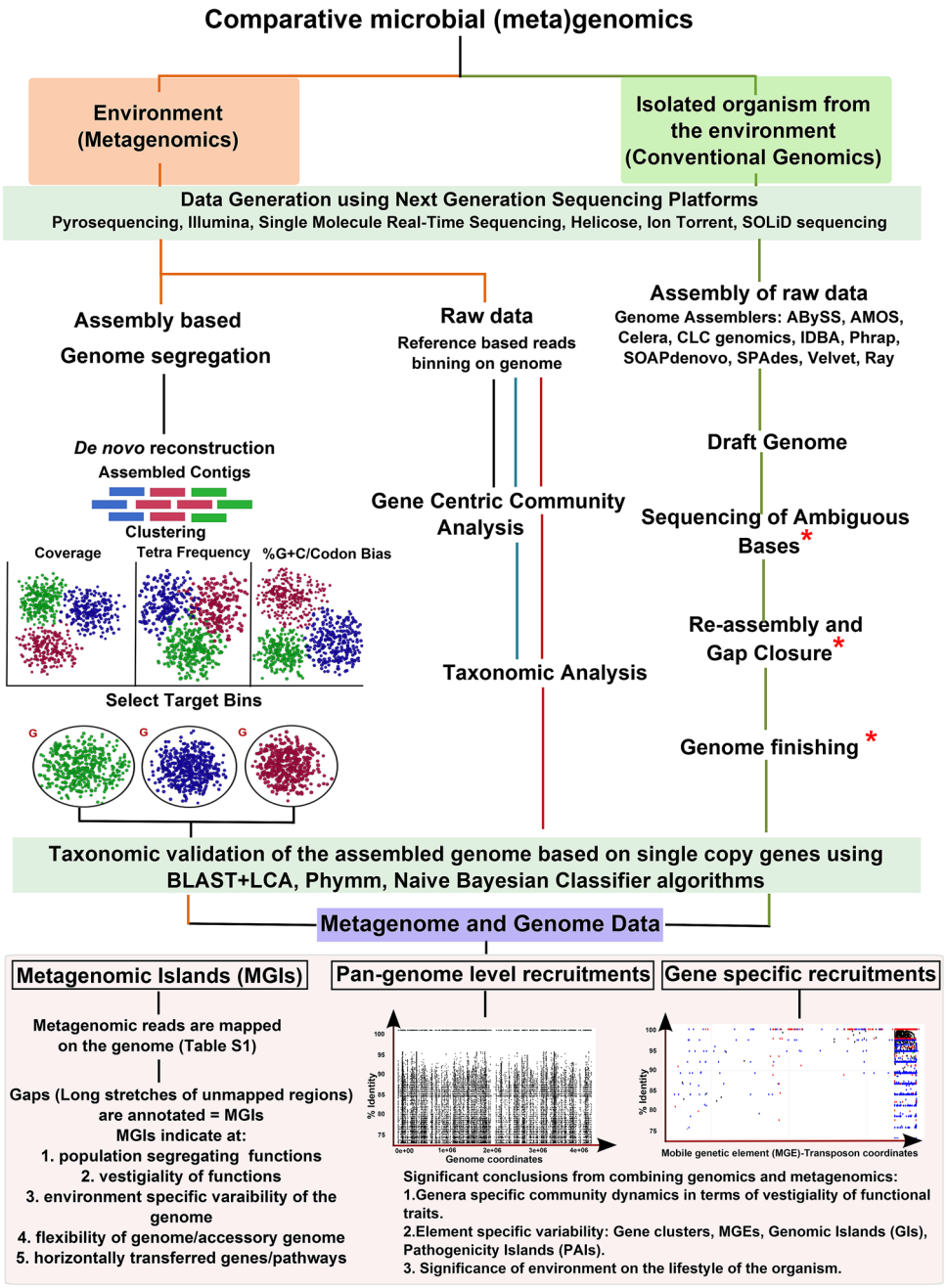

Fig. 1 Schematic representation of workflow for conducting genomic and metagenomic surveys both independently and in association to elucidate community dynamics. The steps with red asterisks are optional steps in the course of analysis. 'G' labeled near bins represent the genomic segregation from metagenome data. The name of assemblers and taxonomic validation techniques mentioned in this figure are mere examples representing generalized methodologies used in the field and does not reflect any biased opinion. However, these methodologies have been reviewed in references Oulas et al. [114] and Sangwan et al. [24]. (The figure was originally produced for this review. The examples for metagenomic recruitments were re-produced from Sharma et al. [17])

gene acquisitions and functional contributions of taxa enriched in an environment [15, 16]. (Meta)genomics can also predict in situ growth dynamics, environment specific lifestyles, seasonal dynamics, and gene-specific genome-wide sweeps across resident populations [17, 18] (Figs. 1, 2). This review intricately covers the (meta)genomic approaches developed across the decade, starting from the metagenome-enabled discovery of 'rare biosphere' [13] until today and how using genomics and metagenomics data conjointly can increase the resolution of investigations concerned with progressive ecological interactions.

(Meta)genomics Enabled Assessment of Population Splitting Factors

When an organism is known to be highly abundant or is isolated from a specific environment along with availability of metagenome database for the same environment, it becomes feasible to recruit metagenome reads over reference genome [2, 19]. This can then be followed by annotation of the unmapped/under-recruited regions known as metagenomic islands (MGIs) [17, 20]. The MGIs are implicated to be mobile genetic elements (MGEs) which

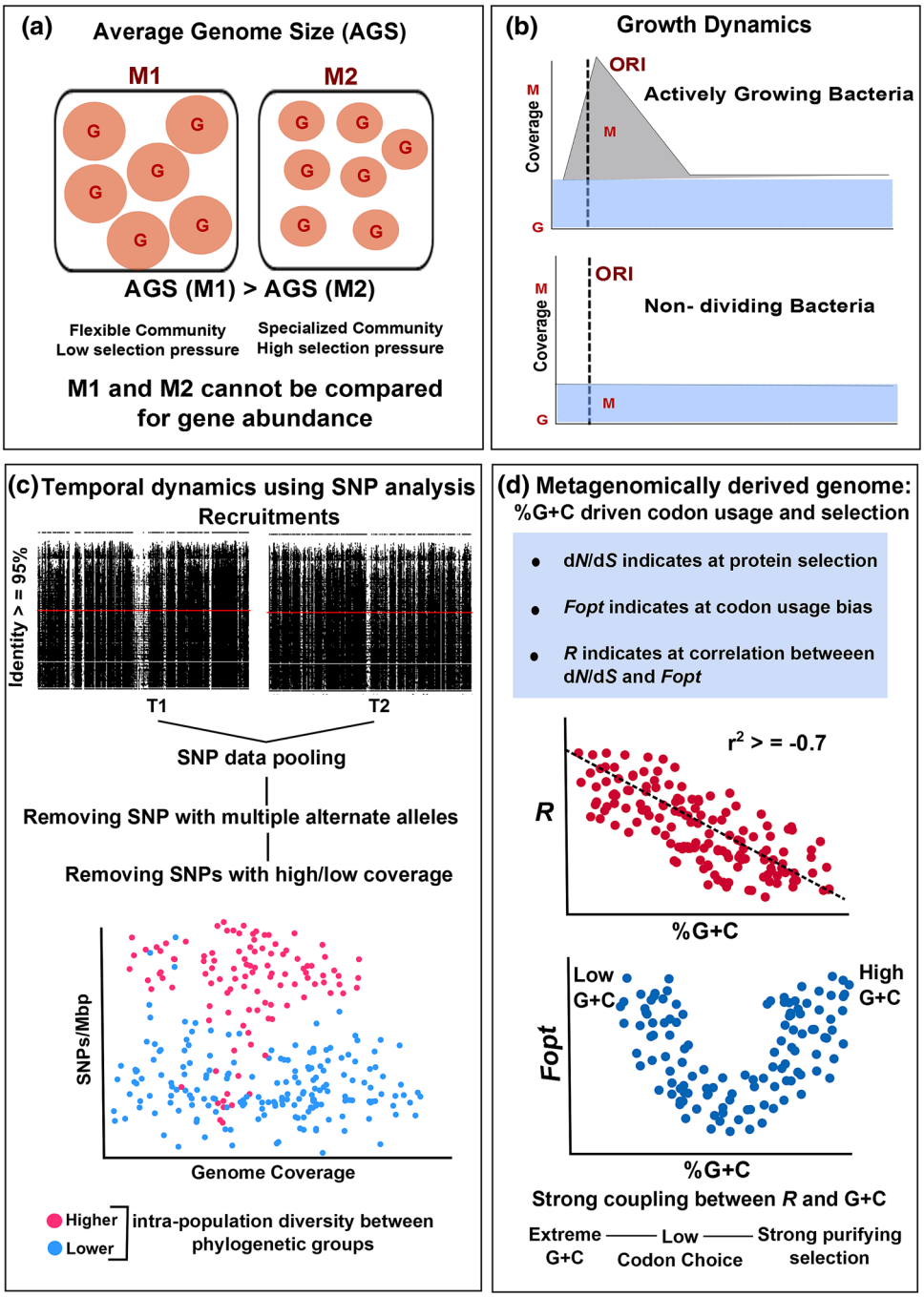

Fig. 2 Diagram showing four major aspects connecting metagenomics (M) and genomics (G) as discussed in the review; **a** average genome size estimation of metagenome data to understand genome selection pressure, **b** understanding the growth dynamics of specific bacterium from metagenome data, **c** SNP analysis using temporal metagenomics, and **d** G+C driven codon usage and protein selection analysis across metagenomic-derived genomes (The figure was originally produced for this review. The concepts for **a**–**d** were taken from Nayfach and Pollard [56], Korem et al. [16], Bendall et al. [15], and Ran et al. [74], respectively)

are part of the accessory bacterial genomes representing highly variable region among different lineages in the population [19]. The steps that can be followed and the outcomes are enumerated in Fig. 1. This analysis as depicted in Fig. 1 becomes more attractive for extreme environments which are characterized by enrichment of the dominant taxa (Fig. 1). For instance, chemical contaminated environment such as hexachlorocyclohexane (HCH) dumpsite has been found to be dominated by Sphingomonads and Pseudomonads [11, 12]. Similarly, soil microcosms are dominantly characterized by *Rhodanobacter*, *Burkholderia*, *Acidobacteria* [13]. The unmapped stretches across the reference genome called as Metagenomic Islands (MGIs) are annotated to highlight the population segregating functions of the in situ cohorts of a specific environment which otherwise is not possible using traditional genomics approach [20] (Table 1). The metagenomic recruitment analyses using environmental data from stressed niches also reveal the population dynamics due to slight environmental perturbations and the

Table 1 List of metagenomic surveys performed in association with the genome reconstruction and reference based recruitments demonstrating different results

S. No.	Environment	Assembly method	Organism	Inference from the study	References
1.	Saturated brines	Metagenomic recruitment	*Salinibacter ruber*	MGI annotations concluded existence of a global strategy of bacteria to escape phage predation across saline environment	[19]
2.	Gut microbiome	Metagenomic recruitment	*Shigella*, *Escherichia coli*, *Neisseria*	Determination of virulence markers	[22]
3.	Freshwater and hypersaline coastal lagoons	Metagenomic recruitment	*Alphaproteobacterium* HIMB114, *Candidatus pelagibacter*, *Synechococcus*	Discovery of novel bacteria, environment specific phylogenetic variations	[1]
4.	Hexachlorocycloh–exane (HCH) dumpsite	Metagenomic recruitment	*Sphingobium*	Horizontal gene transfer events	[12]
5.	Mediterranean sea	Metagenomic recruitment	*Acidimicrobiales*	Differential genotypic enrichment between deep and shallow waters	[2]
6.	Hexachlorocyclohexane (HCH) dumpsite	Metagenomic recruitment	*Pseudomonas*	Integron driven gene transfers	[17]
7.	Hot spring biofilm	Metagenomic recruitment	*Cellulosimicrobium*	Population segregation factor	[20]
8.	Soil	De novo	*Leifsonia*, *Rhodanobacter*, *Acidobacteria*, *Sporolactobacillus*, *Burkholderia*	Deciphering the functional pools of soil ecotypes	[13]
9.	Brackish waters of Baltic Sea	De novo	*Actinobacteria*, *Bacteroidetes*, *Cyanobacteria*, *Verrucomicrobia*, *Alpha-*, *Beta-* and *Gammaproteobacteria* and *Thaumarchaeota*	Seasonal dynamics, lineage-specific variations in terms of gene content	[18]
10.	Gut microbiome	De novo	*Citrobacter*	Growth dynamics using PTR coverage ratio	[16]
11.	Hot spring biofilm	De novo	*Bdellovibrio*	Gene loss	[23]
12.	Brackish waters of Caspian Sea	De novo	*Actinobacteria*, *Thaumarchaea*, and *Alphaproteobacteria*	Phylogenetic placement revealed freshwater or marine origin	[14]
13.	River system	De novo	*Polynucleobacter*	Differential functional constraints	[24]
14.	Freshwater Lake	De novo and metagenomic recruitment	Genomes	Intra-population genetic heterogeneity by SNP analysis over different time points (2005–2013), patterns of gene-gain/gene-loss	[15]

habitat specific genomic alterations over indigenous populations [21].

The recruitment of metagenome reads over reference genomes has been used extensively to demonstrate environment induced variations across genomes fundamentally involving alignment of reads on to microbial whole genomes using global alignment algorithms (Table 1). There are numerous algorithms that can be employed for metagenomic recruitments; few of the most used software/pipelines are discussed in the section below. The MGI annotation across genome of *Salinibacter ruber* using environmental data from saturated brines revealed genetic predominance of the cell wall biogenesis genes across the island region [19]. This suggested that population varied with respect to cell envelopes in saline environment indicating at a global strategy of population against phage predation owing to low eukaryotic grazing pressure [19]. A similar study was performed for identification of pathogenicity markers that employed MGI annotation by recruiting metagenome reads from healthy patients on pathogenic bacteria [22]. This led to mapping of virulence genes specifically in the species with uncharacterized pathogenicity markers such as *Shigella*, *Escherichia coli*, etc. (Table 1). A study using metagenome recruitment data across hexachlorocyclohexane (HCH) dumpsite led to reconstruction of last common ancestor of HCH degrading *Sphingobium* species after discounting MGIs and genomic islands (GIs) which provided an evidence for horizontal gene transfers (HGT) driven acquisition of HCH degrading enzyme arsenal mobilized by environmental pollution [12]

(Table 1). In addition, metagenome reads based recruitment analysis also revealed the pivotal role of integron and its associated transposase gene (TnpA*6100*) in enabling HGTs as a stress response across *Pseudomonas* community inhabiting HCH contaminated environment [17]. Recently, temporal (meta)genomics using recruitments demonstrated intra-population heterogeneity across closely related population of *Methylotenera* via SNP analysis, unraveling the patterns of gene-gain/loss over time highlighting the role of environment in genome modulations [15] (Fig. 2; Table 1).

Tools for Read-Based Metagenomic Recruitments

Metagenomic recruitment largely depends on alignment algorithms and there exist multiple aligners for recruiting/binning metagenomic reads to a reference genome with different execution time and efficiency [25]. The most common formats of raw reads obtained from different sequencing platforms are FASTQ and FASTA. While BLAST and BLAT are the most common alignment algorithms, these are too slow for processing millions of reads generated from a metagenome [26]. Fast alignment approaches include Mapping and Assembly with Qualities (MAQ) [27], Short Oligonucleotide Analysis Package (SOAPv2) [28], Bowtie [29], Basic Oligonucleotide Alignment Tool (BOAT) [30], and Burrows Wheeler Aligner (BWA) [31] (Table 2). Bowtie and BWA both are based on Burrows Wheeler Transform for compressing the reference and the query, making it faster in comparison to BLAST like programs [31]. While MAQ [27] is based on spaced indexing, SOAP uses seed and hash look-up table for query and reference sequence. Between Bowtie and BWA, Bowtie is extensively used for very short reads while BWA is used for mapping of low-divergent sequences on to a large reference genome sequence [26]. Global Alignment Short Sequence Search Tool (GASSST) [32] is software based on global alignments of both short and long reads against large reference sequence with an impeccable edge in its ability to perform fast gapped alignments (Table 2). FR-HIT [26] also performs fragment recruitment with a higher tolerance for mismatches and gaps in contrast to SOAP [28], BWA [31] and Bowtie [29] (Table 2). Similarly, there are many other alignment based algorithms such as MEGAN [33], MetaPhlAn [34], PhymmBL [35], and Kraken [36] (Table 2). MEGAN employs BLAST-based database searching and recruiting the lowest common ancestor (LCA) to the sequence (reads) while PhymmBL uses Markov Model on BLAST results to increase the precision of recruitments [35]. MetaPhlAn assigns taxonomic assignments to sequences by recruiting only a subset of reads to clade-specific markers instead of whole genomes which makes it faster in contrast to other algorithms for huge metagenome datasets [34]. Kraken, which is the fastest of all, uses alignments of *k*-mers over entire microbial genomes achieving relatively higher precision (99.20%) at genus level [36]. All these software/pipelines have been used extensively in numerous studies; however the precision and efficiency might vary from data to data (Table 2).

One of the major challenges in metagenomics is binning of microbial community using very short read sequences. Most of the mapping techniques as discussed above depend on 16S rRNA gene databases or essential genes requiring a read length on a higher side. Most recently Freitas et al. [37] used hierarchical array of unique signatures. Current taxonomic profiling methodologies stay biased as gene based approach depends heavily on correct coding orientations which is not achievable while analyzing metagenome short reads data. GOTTCHA pipeline uses machine learning to determine the unique genomic region followed by deciphering the distribution and coverage of these specific regions [37]. Hence, depending on the type of raw data and system configuration available at our end, we can decide on the software to be used for metagenomic recruitment (Table 2).

Metagenomic recruitments of reference genomes highlight the relative abundance of genomes or pan-genomes at the sampling site by using 80–95% of identity cut-off for alignments, accommodating for species level recruitments [2, 12, 19] (Fig. 1). While exploring the population dynamics of a particular strain, stringency of the identity cut-off can be increase up to 98–100%. For (meta)genome recruitments generally the % identity threshold is defined as the number of identities between read and reference divided by the average read length. This value has been standardized as 80% for metagenomic binning over genome i.e. requiring 80% identity over 80% length of the read [19]. Metagenome tilling can thus highlight the modulation of the accessory genomes delineating the population splitting factors across bacterial ecotypes by annotating the MGIs (Fig. 1). However, there still lies a bottleneck of producing false-positives because of sequencing bias. Therefore, manual curation remains the most important step including scanning for tRNA flanking the island regions, differential G+C content, tetranucleotide frequency, and codon usage skew [19]. The downstream analysis becomes very important to confirm the accuracy of MGIs across a genome after recruitment of metagenome reads on to reference genomes. Annotation of the unmapped regions is achieved by basic database search using programs such as BLAST [38], GHOSTX [39], GHOSTZ [40], BLAT [41], HMMER [42] etc. against databases such as NCBI-nr [43], KEGG [44], Pfam [45], SwissProt [46] etc.

Table 2 List of software for recruitment of metagenomic reads

S. No.	Software	Size of the reads mapped	Operating system	Language written in	Algorithm	Limitations	Running time[a]	References
1.	BLAT	200,000 bp	Windows, Linux, Web-based	C++	Seed and extend	Cannot handle database larger than 4 Gb	~0.002 s/read	[41]
2.	RMAP	Up to 500 bp	Unix/Linux, Mac OS X, POSIX OS	C++	Seed and extend with increased mapping mismatch scores	Cannot handle insertions/deletions	~0.460 s/read	[48]
3.	BOAT	Up to 150 bp	Unix like-Linux, Solaris	C++	Index and search	High computation time, cannot support paired-end alignment for repetitive region mapping	~0.120 s/read	[30]
4.	SOAPv2	Up to 1024 bp	Linux, Mac OS X	C++	Seed and hash look-up table	Uses large memory i.e. >5 Gb for processing	~0.001 s/read	[28]
5.	Bowtie	Up to 1024 bp	Linux, Windows, Mac OS X	C++	Burrows wheeler transform	Relatively lesser confident mapping	~0.001 s/read	[29]
6.	GASSST	36–500 bp	Linux	C++	Seed and extend	Cannot support paired-end alignment for repetitive region mapping	~1.39e−5 s/read	[32]
7.	FR-HIT	Up to 3000 bp	Linux	C++	*k*-mer based	Generates false positives across conserved sequences	~0.005 s/read	[26]
8.	Kraken	Up to 200 bp	Linux	C++, Perl	*k*-mer based	Relatively low precision (*F*-score)	0.04e−2 s/read	[36]
9.	Genomic Origins Through Taxonomic CHAllenge (GOTTCHA)	Starting from 100 bp	Linux, Mac OS X	C, C++	Unique Signature based (not gene based)	Misclassifications at phylum level	~0.001 s/read	[37]
10.	Naïve Bayes Classifier (NBC)	Up to ~ 250 bp	Linux, Windows, Web Server	C++, Java	Naïve Bayes Algorithm-machine learning	Not suitable for longer reads	~9 s/read	[50]
11.	PhymmBL	Starting from 100-bp	Linux, Windows	C++, Perl	Markov Model	False positives and slow	~0.630 s/read	[35]
12.	BWA	70–100 bp	Linux	C++, Java, Perl, C	Burrows wheeler transform	Error rate is high in case of longer reads	~0.001 s/read	[41]
13.	MAQ	Up to 63 bp	Linux	C, C++, Perl	Index, extend and score	Cannot support single-ends read mapping and longer reads	~0.017 s/read	[27]
14.	MUMmer 3.0	Doesn't depend on the size of reads	Windows, Linux, Mac OS X	Java	Suffix trees generation for anchor finding	Limits the size of reads file that can be recruited	~17 s/5 Mb of reference genome	[47]

Table 2 continued

S. No.	Software	Size of the reads mapped	Operating system	Language written in	Algorithm	Limitations	Running time[a]	References
15.	Megan	35–800 bp	Windows, Linux, Mac OS	Java	Lowest common ancestor (LCA)	Higher non-specific recruitments	~ 2.25 s/read	[33]
16.	MrFAST	>25 bp (up to 35 bp)	Unix/Linux	C	Seed and extend	Not suited for INDEL calling	~ 0.009 s/read	[49]
17.	MetaPhlAn	<400 bp	Linux	Python	Clade-specific markers based	Recruits only a subset of reads specific to marker genes rather than whole genome	~ 0.001 s/read	[34]

[a] The running time is not based on identical RAM configuration

De novo Segregation of Metagenome Datasets into Genome Bins

While metagenomic recruitments on a genome can provide insights into environment specific genome modulations, assembly of near complete genomes using metagenome binning can provide accurate functional contribution of an individual genotype/population in a complex community. Nonetheless achieving a high fidelity bin without any cross contamination at strain/species level resolution remains a challenge with a moderately (coverage) sequenced metagenome (Fig. 1). The coverage (sequencing depth) of a sequenced metagenome remains the most significant parameter while recovering a genome from environmental data [51]. However, the ever expanding field of NGS overcomes this bottleneck leading to the assembly of genomes even with <1% relative abundance in a metagenome [23].

Essentially the recovery of near complete genomes from metagenomes is based on alignments against reference genomes and reference databases remain limited due to an overwhelming unexplored complex community exceeding the limited reference databases [52, 53]. De novo metagenome segregation approach uses tetranucleotide frequencies, G+C skew, and coverage which are assumed to be relatively constant across one genome (Fig. 1). However, there are known genomes having inconsistent base compositions and G+C content which compromise this de novo assembly approach [53]. Another method is based on retrieving set of specific genes such as 107 essential genes [54], 31 bacterial marker genes [55] directly from metagenome to separate organisms. These methodologies although extremely used are based on gene abundance which is sometimes exactly identical for closely related organisms and therefore known to be co-abundant [52]. Recently 7381 co-abundance gene groups (CAGs) were used to recover genomes from 396 human gut microbiome samples [52]. This methodology was used to assemble 238 microbial genomes belonging to archaea, bacteria and viruses. In order to perform benchmarking, 19 of the sampled individuals were fed with fermented milk products containing *Bifidobacterium animalis* subsp. *lactis* CNCM I-2494 which also has been already sequenced. Using CAGs *B. animalis* genomes was reconstructed and 95% of *B. animalis* reference genes were recovered from the genomes with 99.9% identity with respect to the reference *B. animalis* subsp. *lactis* CNCM I-2494 [52]. Hence, it is suggested to use co-abundance gene profiles of environmental metagenomes which is capable of segregating taxonomically related microorganisms with a higher accuracy in contrast to gene-based or composition based approaches.

Significance of Average Genome Size (AGS) Estimations Across Metagenome

Metagenomics besides being used for community profiling, can also be used to determine relative abundance of gene families and pathways between different sites. In order to accurately detect the gene abundance, it is most important to determine the average genome size (AGS) to give a statistically significant interpretation of variations of gene abundances between different metagenomes [56, 57]. AGS can be simply explained as the average of sizes of genomes present in a metagenome, which can vary between two sites thus introducing gene frequency variations or unreal variations [58]. Therefore, while comparing metagenomes with different AGS, false positives can be observed or sometimes stability among genes can be demonstrated between sites even when there is difference in real (i.e. false negatives) (Fig. 2a) [59]. AGS also is significant in estimating the evolutionary forces active on an organism thriving in a particular environment. Bacterial genome size demonstrates the environmental pressures, community, metabolic preferences and lifestyle of an organism [60]. For instance, it has been observed that bacteria with relatively larger genome sizes follow a generalist lifestyle whereas bacteria with smaller genome sizes follow more specialized lifestyle [61].

There are multiple softwares to estimate AGS across a metagenome including most used Genome relative Abundance and Average Size (GAAS) [60] which is based on BLAST searches of the metagenome reads data against a database of microbial genomes including bacteria, archaea and viruses. This approach stays biased as there are microorganisms which are not submitted in the database and the metagenome sample to be analyzed might contain novel organisms. Hence, GAAS is not a choice in case we are analyzing a metagenome of a unique niche. However, Raes et al. [58] had devised an approach where the AGS was calculated based on the abundance of reads assigned to 35 essential genes, which made it much faster. But this approach still carried a limitation i.e. it could only analyze metagenomic reads of length greater than 300 bp. In case of newer sequencing platforms where short reads library preparation is used, this methodology can't be used with accuracy. In order to calculate AGS accurately, by overcoming above mentioned problems, recently Nayfach and Pollard [56], designed a pipeline called "MicrobeCensus". MicrobeCensus depends on determining reads density on the housekeeping genes and can also be applied on to metagenome reads as short as 50 bp. This software considers 40 marker genes for the domains of bacteria and archaea and 114 for all bacteria only [62]. The markers list for 40 genes comprises of ribosomal protein units S2, S10, L1, L22, L4, L2, S9, L3, L14B, S5, S19, S7, L16, S13, L15, L25, L6, L11, L5, S12, L29, S3, S11, L10, S8, L18P, S15P, S17, L13, L24, translation elongation factor EF-2, translation initiation factor IF-2, metalloendopeptidase, ffh signal recognition particle protein, phenylalanyl-tRNA synthetase beta subunit, phenylalanyl-tRNA synthetase α subunit, tRNA pseudouridine synthase B, porphobilinogen deaminase, phosphoribosylformylglycinamidine cyclo-ligase, and ribonuclease HII [62]. Further, the accuracy of MicrobeCencus was established using "Median Unsigned Error" in order to account for errors due to over- and under-representation of sequences.

Using (Meta)genomics Data in Deciphering Growth Dynamics of Bacteria

The metagenomic sequence data can provide an understanding of the presence of microbiota at a particular niche including gut, biofilm, hot spring, etc. Recently Korem et al. [16] devised a way to obtain information on growth dynamics of particular bacteria/genome enriched in a metagenome. This methodology focuses on examining pattern of sequencing coverage specifically at origin of replication of bacterial genomes (Fig. 2b). It is well established that bacterial replication initiates at origin of replication (ori) bidirectionally, hence the regions already replicated will have two copies of ori in contrast to unreplicated regions. The same concept was earlier applied only on yeast cells with coordinated stage of replication [63]; however it stands true for all the cells at any stage of replication as well [64, 65]. Using genome data from multiple bacteria, it was found that the region in the proximity of ori is present in high copy number in actively growing bacteria as compared to the DNA segment present towards the terminus [63] (Fig. 2b). The ratio of copy number of region near ori to the DNA segment near terminus is termed as peak-to-trough ratio (PTR) and is a direct measure of growth rate of the bacterial genome [66]. PTR ratio of greater than 1:1 is a quantitative indicative of higher growth dynamics. This quantitative relationship was further proven experimentally using *E. coli* (strain K-12) cultures [16]. Similar patterns were observed across *E. coli* genomes retrieved from human fecal metagenome samples to that of in vitro cultures [52, 67, 68]. Using *E. coli* genomes from 583 databases, it was found that PTR varied from 1 to 2.4 which was similar to the ratio obtained in in vitro experiments i.e. 1–2.6 [16].

Further, as an extension of this concept, it was found that PTRs can also monitor clinical changes after treatment by antibiotics. For this, *Citrobacter rodentium* was treated with antibiotic erythromycin and PTRs reduced drastically.

The reduction was evident within 30 min after administration. However, during antibiotic recovery (washing of cultures) an increase was observed. In order to determine the virulent bacterial activity in a disease condition, *C. rodentium* strains (both virulent and non-virulent) were observed for PTR patterns. For the first five days both the virulent and non-virulent strains showed similar PTR values, nevertheless, at 6–9 days post infection PTR values for virulent strains increased drastically in comparison to non-virulent strains. This was justified as an indication of mucosal adhesion and proliferation by virulent strains in contrast to non-virulent strains at that stage [16]. Furthermore, the FDR (corrected *P* value <0.005) association showed significant correlations between PTRs and several disorders/metabolic conditions like Crohn's Disease [52], ulcerative colitis [69], fasting serum insulin, fasting blood glucose, and Type II diabetes [68].

Evidence for Genome-Wide Sweeps Using Temporal (Meta)genomics

Microbial communities are comprised of distinct phylogenetic groups within ecologically coherent populations due to high recombination rates of superior genes between the members of population [70]. In order to study the genetic heterogeneity, time-series metagenomics along with *de novo* genome assembly holds a great potential by directly tracking the evolutionary patterns driven gene-gain/loss throughout [15]. De novo reconstruction of genomes from the metagenome data provide the reference genomes which are again recruited over by the metagenome reads for SNPs deciphering the genetic diversification within discrete populations [71] (Fig. 2c). This has further enabled us to directly encapsulate evolutionary models such as genome-wide sweeps, which were not studied earlier [72] in the same yet phylogenetically diverse environment.

Metagenomic recruitment over the assembled reference genomes demonstrate two types of populations; one called 'sequence discrete' populations with recruitment at ≥99% identity and the other called 'close sympatric' populations with recruitment <90%. Sequence discrete populations represent highly similar genotypes with low extent of diversity which can be analyzed for SNPs. Bendall et al. [15] reported significantly different i.e. eightfold (SNPs per Mbp) for two close genotypes of the same genus i.e. *Methylotenera* assembled from 9-year study period indicating at astounding intra-population diversity (Fig. 2c). However, most of the SNPs for discrete populations did not show amino-acid substitutions. Further, time series metagenomics can also unveil the gene-gain/loss patterns across one specific population (reconstructed genome) [15]. In case the relative abundance for a specific set of genes is increased over time, it suggests that the gene was acquired horizontally, whereas in the case of decrease in frequency of genes proposes that the newly dominant lineage will eventually lack these genes for a specific population (i.e. genus or phylum or order) given a constant functional constraint. Hence, genome wide studies using temporal metagenomics can provide a clear understanding of both genome-wide sweeps and gene-specific sweeps taking place across intra- and inter-populations which can further explore evolutionary models controlling population dynamics of an environment.

Using population Genomes to Analyze Differential Codon Usage Preferences

It has been recently established that microbial communities at extreme environments evolve faster characterized by a strong purifying selection in order to undergo genome optimization under specific functional constraints [73]. De novo genome reconstruction of uncultivable diversity from extreme metagenomes can be used to explore relationship between %G+C, codon usage preferences and protein selection to validate the evolutionary pressures acting on the bacterial community under stressed environments (Fig. 2d). In order to derive the coupling force, correlation (*R*) between gene-specific optimal codon frequencies (F_{opt}) (an indicative of codon bias) and d*N*/d*S* is calculated [74]. It has already been established that strongest positive coupling exists across *Cyanobacteria* and *Tenericutes* followed by *Firmicutes*, *Spirochaetes* etc. whereas, *Actinobacteria* group has strong but negative correlation. Generally, low negative *R* highlights the selection of "high-status" genes which are central to metabolic pathways and thus evolve slowly with overwhelming purifying selection pressure (Fig. 2d). The value of *R* also shows significant relationship with %G+C. For genomes characterized with extreme %G+C, the *Fopt* values tend to be on higher side (Fig. 2d). Thus, the dependence of selection pressure in terms of codon preferences on nucleotide composition can provide insights into the poorly understood evolution patterns. Therefore, using metagenomics based genome reconstruction; the habitat specific evolutionary pressure can be estimated employing the genome data.

Interestingly, it was established that codon usage skew is specific to metagenome as a whole and is independent of the bacterial community enriched in the metagenome data [75]. This suggested that bacterial genera in the same metagenome can exhibit variable codon usage preferences but overall the metagenome is characterized by an accumulative codon bias which differs from the other

metagenome samples markedly just like an observation for single microbial species/genome [76, 77]. To investigate phyletic independence, the species specific genes common between different metagenome samples were retrieved and distances were calculated between codon usage preferences of each [78]. It was found that codon usage of compared phylogenies showed greater variations between metagenomes than in different species of the same metagenome. Similar analysis has also been extended to variable environmental conditions which demonstrated that different species show lower variability of codon usage in case of constrained environmental conditions [79–81]. Further, segregated genomic data from the metagenome showed consistent codon usage patterns within a genome of a metagenome. Hence, the constitution of sequence composition across genome and metagenome can elucidate evolutionary pressures in terms of codon choices and protein selection.

Application of (Meta)genomics in Clinical Microbiology

HGT driven variations are main contributors of transition of non-pathogenic bacteria into pathogenic bacteria and vice versa. Perna et al. [82] compared pathogenic *E. coli* O157:H7 to the non-pathogenic bacteria *E. coli* K-12 which led to the identification of candidate genes specifically responsible for pathogenesis of the pathogenic *E. coli* strains [82]. Furthermore, comparative genomics analysis across *Bacillus strains* revealed the significance of mobile genetic elements in imparting virulence to bacterial strains [83]. Genomic analyses between pathogenic and non-pathogenic *Mycobacterium tuberculosis* strains revealed new set of pathways in pathogenic strains in contrast to avirulent strains [84]. Most significant finding of this study was discovery of alternate metabolic pathways which shed light into their mechanisms of pathogenesis thus providing a base for developing diagnostic markers against tuberculosis [84]. Comparative genomics has provided significant information regarding credible virulence determining factors that can be further targeted for vaccine development. While comparative genomics has provided insights into the bacterial evolution of pathogenesis, the metagenomics approach has also been used for functional screening of virulence markers overall at an environment [85]. Metagenomics approaches have focused on the predominance of pathogenic bacteria in natural environments such as human gut. For instance, Sommer et al.[86] used metagenomics data to characterize antibiotic resistance potential of healthy human microbiome. A very strong correlation has been established between human microbiota imbalance and diseases such as irritable bowel disease [87], obesity [88], cystic fibrosis [89] etc. A metagenomics approach along with a functional screening of potential pathogenicity markers and antibiotic resistance has been used to investigate complex environments [85]. Metagenomic studies have also demonstrated the effect of factors such as environment, geographical location, antibiotics, age, and diet on the human gut ecosystem. But the metagenome sequencing not only sequences the pathogenic sequences but can also capture the human genetic sequences [90]. This might lead to an incorrect understanding of the pathogenic community hosted by human body. However, this has an advantage of providing access to the genetic changes that might be taking place in human body under diseased condition [91]. Additionally, shotgun sequencing can also identify pathogenic species at strain level resolution since it is based on whole genome based markers rather than only 16S rRNA gene [92]. This has been already reported in the metagenomic sequencing of cholera patients [93], tuberculosis [94], *E. coli* [95], and methicillin-resistant *Staphylococcus aureus* (MRSA) [96].

Another aspect of metagenomics in the field of clinical microbiology is targeted antimicrobial therapy after accurate diagnosis of the pathogen which can reduce antibiotic-associated side effects due to broad-spectrum antibiotic regimen [97]. This has led to a significant reduction of mortality rates in patients of ventilator associated pneumonia (VAP) [98]. Metagenomics-enabled accurate and rapid diagnosis of infectious diseases along with understanding of antibiotic resistance pattern can empower the physicians to use targeted antimicrobial therapies [99]. Therefore, both metagenomics and single-species-targeting genomics of pathogenic bacteria can provide insights into the pathogenesis and a better understanding of the virulence markers providing a platform for pathogenic diagnostics [100].

Application of (Meta)genomics in Fecal Microbial Transplants

Microbiome analysis is the most recent extension of metagenomics as of today which makes metagenomics a direct application in human health. The existence of symbiotic relationship between gut microbiota and human health is well established and human intestine is known to harbor around 10^{14} microbes with 35,000 different species [101, 102]. Interestingly, the number of microbes is nearly 10 times more than the number of cells in the human body [103]. Human gut microbiota is known to play significant roles in postnatal structural and functional maturation of gut, development of immune system and nervous system [104–106]. Gut microbiota is also identified as to produce antimicrobial proteins such as cathelicidins, defensins and

C-type lectins [107, 108]. Imbalance of microbiota can lead to disease states such as antibiotic-associated diarrhea and *Clostridium difficile* infection (CDI). Fecal microbiota transplantation has proven to be very helpful by restoring the disturbed microbiota. This was first reported by Ge Hong, who used fecal transplantation in treating food poisoning [109], however in modern medicine it was used for the first time to treat pseudomembranous colitis [110]. As of today there are many clinical reports on using FMT for disease conditions like CDI, autism, depression, inflammatory bowel disease, Parkinson, multiple schlerosis, obesity [111]. In this scenario metagenomics play a significant role by determining the microbial content in both healthy and diseased gut before and after the transplant. In addition to this metagenomics identify dysbiosis/imbalance of the human gut microbiome in the diseases, and can also determine novel changes in microbial functions [112].

Computational Challenges in Data Interpretation

With the advent of NGS, data generated for genomes or metagenomes include millions of short reads, which before any downstream analysis need to be assembled into manageable data (i.e. genome/metagenome) [113]. Multiple state-of-art assemblers such as Velvet, Ray, ABySS can assemble gigabytes of data into 10 and 1000 s of contigs of genome and metagenome, respectively [113]. Broadly, there are two types of assemblers: (1) reference based and (2) de novo assemblers [114]. Reference based assemblers can be used when there is availability of reference genomes to be used to order the contigs. These include MIRA4, MetaAMOS, Newbler which are not computationally exhaustive and use a closely related reference genome already deposited in databases [115, 116]. This set of assemblers however remains biased due to limitation of existing databases and cannot be used while exploring a unique environment [113]. De novo assemblers can assemble the raw reads into contigs based on graph theories like de-Bruijn graph without any reference genome [117]. Tools such as Velvet, MetaVelvet, ABySS, SOAP, SPAdes, Ray Meta, Meta-IDBA etc. are among the most used softwares as of today [28, 118–123]. Due to processing of multiple nodes during assembly of reads, the de novo assemblers are computationally quiet extensive yet best suited while exploring unique environments. Metagenome assembly has improved over time but still it carries many challenges as of today majorly due to computational memory constraints and the biological complexity of the data [53]. The population bias introduced due to sequencing leads to predominance of specific genomes and no or less coverage for others [124]. Hence, in order to correctly assemble the data, coverage of reads needs to be on a higher side ($\sim 10\times$) [124]. The sequencing errors such as repeats incorporation has also been challenging for assembly as they can be misinterpreted for identical regions in one genome or conserved regions across different species or homologous segments across closely related strains [125, 126]. Under these circumstances precise analyses of assembly metrics such as N50, average coverage, and total assembly size can be used to measure the efficiency of good assembly [127]. Detailed discussion of (meta)genome assemblers remain outside the scope of this review article: for details please review Refs. [53, 114].

We have now entered the era of challenged data-interpretation shifting from the era of restricted data-generation. With consistently increasing data, there is a need for algorithms which can compare huge amount of data. Multiple algorithms are being scripted every day; however, they need large memory and specific hardware options which can be challenging. In addition, every goal in (meta)genomics requires a different set of algorithms for a specific objective. There is an increased improvement in development of data visualization tools given the significance of visualization of data in complete data analysis [128]. Genome analyses tools are largely command line based and does not work using Graphic User Interface (GUI) very efficiently due to high throughput data which hinders the progress of biology labs in this field and encourages collaborations across multidisciplinary labs.

Conclusions

Expansion of the emerging fields of genomics and metagenomics can provide an access to the complete genetic content of a bacterium of interest and community profile of an environment, respectively. However, the conventional study of genomics needs culture-based bacterial isolation which offers huge bias since more than 99% of the microorganisms are uncultivable. Therefore, metagenomics not only provides an overall taxonomic composition of an environment exhibiting the presence or absence of microbial entities but can also target a single unculturable bacterial (species/strain) genomics surpassing the need for isolation. This review comprehensively surveys the most recent techniques using both genomics and metagenomics data together which can provide detailed insights into environmental microbiology (Fig. 2).

Acknowledgements The authors acknowledge funds from Government of India under project from Department of Biotechnology (DBT), National Bureau of Agriculturally Important Microorganisms (NBAIM) AMASS/2006-07/NBAIM/CIR, University of Delhi Research & Development (R&D) grant 2015-16, and DU-DST Promotion of University Research and Scientific Excellence (PURSE). AS gratefully acknowledge funds from NBAIM.

Author contributions AS and RL fabricated the concept and wrote the review.

Compliance with Ethical Standards

Conflict of interest The authors declare no conflict of interest.

References

1. Ghai R, Hernandez CM, Picazo A, Mizuno CM, Ininbergs K, Díez B, Valas R, DuPont CL, McMahon KD, Camacho A, Rodriguez-Valera F (2012) Metagenomes of Mediterranean coastal lagoons. Sci Rep 2:490. doi:10.1038/srep00490
2. Mizuno CM, Rodriguez-Valera F, Ghai R (2015) Genomes of planktonic *Acidimicrobiales*: widening horizons for marine *Actinobacteria* by metagenomics. mBio 6:e02083-14. doi:10.1128/mBio.02083-14
3. Negi V, Lata P, Sangwan N, Gupta S, Das S, Rao DLN, Lal R (2014) Draft genome sequence of Hexachlorocyclohexane (HCH)-degrading *Sphingobium lucknowense* Strain F2, isolated from the HCH Dumpsite. Genome Announc 2:e00788-14. doi:10.1128/genomeA.00788-14
4. Sharma A, Hira P, Shakarad M, Lal R (2014) Draft genome sequence of *Cellulosimicrobium* sp. MM, isolated from arsenic rich microbial mats of a Himalayan Hot Spring. Genome Announc 5:e01020-14. doi:10.1128/genomeA.01020-14
5. Singh AK, Sangwan N, Sharma A, Gupta V, Khurana JP, Lal R (2013) Draft genome sequence of *Sphingobium quisquiliarum* P25T, a novel Hexachlorocylohexane (HCH) degrading bacterium isolated from the HCH Dumpsite. Genome Announc 1:e00717-12
6. Mukherjee U, Kumar R, Mahato NK, Khurana JP, Lal R (2013) Draft genome sequence of *Sphingobium* sp. HDIPO4, an avid degrader of Hexachlorocyclohexane. Genome Announc 1:e00749-13. doi:10.1128/genomeA.00717-13
7. Kaur J, Verma H, Tripathi C, Khurana JP, Lal R (2013) Draft genome sequence of a Hexachlorocyclohexane-degrading bacterium, *Sphingobium baderi* Strain LL03T. Genome Announc 1:e00751-13. doi:10.1128/genomeA.00751-13
8. Dua A, Sangwan N, Kaur J, Saxena A, Kohli P, Gupta AK, Lal R (2013) Draft genome sequence of *Agrobacterium* sp. Strain UHFBA-218, isolated from rhizosphere soil of crown gall-infected cherry rootstock colt. Genome Announc 1:e00302-13. doi:10.1128/genomeA.00302-13
9. Dua A, Malhotra J, Saxena A, Khan F, Lal R (2013) *Devosia lucknowensis* sp. nov., a bacterium isolated from Hexachlorocyclohexane (HCH) contaminated pond soil. J Microbiol 51:689–694. doi:10.1007/s12275-013-2705-9
10. Dwivedi V, Sangwan N, Nigam A, Garg N, Niharika N, Khurana P, Khurana JP, Lal R (2012) Draft genome sequence of Thermus sp. RL isolated from hot water spring located atop the Himalayan Ranges at Manikaran. India. J Bacteriol 194:3534–3535. doi:10.1128/JB.00604-12
11. Sangwan N, Lata P, Dwivedi V, Singh A, Niharika N, Kaur J, Anand S, Malhotra J, Jindal S, Nigam A, Lal D, Dua A, Saxena A, Garg N, Verma M, Kaur J, Mukherjee U, Gilbert JA, Dowd SE, Raman R, Khurana P, Khurana JP, Lal R (2012) Comparative metagenomic analysis of soil microbial communities across three Hexachlorocyclohexane contamination levels. PLoS ONE 7:e46219. doi:10.1371/journal.pone.0046219
12. Sangwan N, Verma H, Kumar R, Negi V, Lax S, Khurana P, Khurana JP, Gilbert JA, Lal R (2014) Reconstructing an ancestral genotype of two hexachlorocyclohexane-degrading *Sphingobium* species using metagenomic sequence data. ISME J 8:398–408. doi:10.1038/ismej.2013.153
13. Delmont TO, Eren AM, Maccario L, Prestat E, Esen OC, Pelletier E, Le Paslier D, Simonet P, Vogel TM (2015) Reconstructing rare soil microbial genomes using in situ enrichments and metagenomics. Front Microbiol 6:358. doi:10.3389/fmicb.2015.00358
14. Mehrshad M, Amoozegar MA, Ghai R, Fazeli SA, Rodriguez-Valera F (2016) Genome reconstruction from metagenomic datasets reveals novel microbes in the brackish waters of the Caspian Sea. Appl Environ Microbiol 82:1599–1612. doi:10.1128/AEM.03381-15
15. Bendall ML, Stevens SLR, Chan L-K, 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 (2016) Genome-wide selective sweeps and gene-specific sweeps in natural bacterial populations. ISME J 10:1589–1601. doi:10.1038/ismej.2015.241
16. Korem T, Zeevi D, Suez J, Weinberger A, Avnit-Sagi T, Pompan-Lotan M, Matot E, Jona G, Harmelin A, Cohen N, Sirota-Madi A, Thaiss CA, Pevsner-Fischer M, Sorek R, Xavier RJ, Elinav E, Segal E (2015) Growth dynamics of gut microbiota in health and disease inferred from single metagenomic samples. Science 349:1101–1106. doi:10.1126/science.aac4812
17. Sharma A, Sangwan N, Negi V, Kohli P, Khurana JP, Rao DLN, Lal R (2015) Pan-genome dynamics of *Pseudomonas* gene complements enriched across hexachlorocyclohexane dumpsite. BMC Genom 16:1. doi:10.1186/s12864-015-1488-2
18. Hugerth LW, Larsson J, Alneberg J, Lindh MV, Legrand C, Pinhassi J, Andersson AF (2015) Metagenome-assembled genomes uncover a global brackish microbiome. Genome Biol 16:279. doi:10.1186/s13059-015-0834-7
19. Pašić L, Rodriguez-Mueller B, Martin-Cuadrado A-B, Mira A, Rohwer F, Rodriguez-Valera F (2009) Metagenomic islands of hyperhalophiles: the case of *Salinibacter ruber*. BMC Genom 10:1. doi:10.1186/1471-2164-10-570
20. Sharma A, Gilbert JA, Lal R (2016) (Meta)genomic insights into the pathogenome of *Cellulosimicrobium cellulans*. Sci Rep 6:25527. doi:10.1038/srep25527
21. Johnston ER, Rodriguez-R LM, Luo C, Yuan MM, Wu L, He Z, Schuur EA, Luo Y, Tiedje JM, Zhou J, Konstantinidis KT (2016) Metagenomics reveals pervasive bacterial populations and reduced community diversity across the Alaska Tundra ecosystem. Front Microbiol 7:579. doi:10.3389/fmicb.2016.00579
22. Belda-Ferre P, Cabrera-Rubio R, Moya A, Mira A (2011) Mining virulence genes using metagenomics. PLoS ONE 6:e24975. doi:10.1371/journal.pone.0024975
23. Sangwan N, Lambert C, Sharma A, Gupta V, Khurana P, Khurana JP, Sockett RE, Gilbert JA, Lal R (2015) Arsenic rich Himalayan hot spring metagenomics reveal genetically novel predator-prey genotypes. Environ Microbiol Rep 7:812–823. doi:10.1111/1758-2229.12297
24. Sangwan N, Zarraonaindia I, Hampton-Marcell JT, Ssegane H, Eshoo TW, Rijal G, Negri MC, Gilbert JA (2016) Differential functional constraints cause strain-level endemism in *Polynucleobacter* populations. mSystems. doi:10.1128/mSystems.00003-16
25. Rusch DB, Halpern AL, Sutton G, Heidelberg KB, Williamson S, Remington K, Beeson K, Tran B, Smith H, Baden-Tillson H, Stewart C, Thorpe J, Freeman J, Andrews-Pfannkoch C, Venter JE, Li K, Kravitz S, Heidelberg JF, Utterback T, Rogers YH, Falcón LI, Souza V, Bonilla-Rosso G, Eguiarte LE, Karl DM, Sathyendranath S, Platt T, Bermingham E, Gallardo V, Tamayo-Castillo G, Ferrari MR, Strausberg RL, Nealson K, Friedman R, Frazier M, Venter JC (2007) The Sorcerer II global ocean sampling expedition: northwest Atlantic through eastern tropical pacific. PLoS Biol 5:e77. doi:10.1371/journal.pbio.0050077

26. Niu B, Zhu Z, Fu L, Wu S, Li W (2011) FR-HIT, a very fast program to recruit metagenomic reads to homologous reference genomes. Bioinformatics 27:1704–1705. doi:10.1093/bioinformatics/btr252
27. Li H, Ruan J, Durbin R (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res 18:1851–1858. doi:10.1101/gr.078212.108
28. Li R, Li Y, Kristiansen K, Wang J (2008) SOAP: short oligonucleotide alignment program. Bioinformatics 24:713–714. doi:10.1093/bioinformatics/btn025
29. Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10:R25. doi:10.1186/gb-2009-10-3-r25
30. Zhao S-Q, Wang J, Zhang L, Li J-T, Gu X, Gao G, Wei L (2009) BOAT: basic oligonucleotide alignment tool. BMC Genom 10:S2. doi:10.1186/1471-2164-10-S3-S2
31. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 25:1754–1760. doi:10.1093/bioinformatics/btp324
32. Rizk G, Lavenier D (2010) GASSST: global alignment short sequence search tool. Bioinformatics 26:2534–2540. doi:10.1093/bioinformatics/btq485
33. Huson DH, Auch AF, Qi J, Schuster SC (2007) MEGAN analysis of metagenomic data. Genome Res 17:377–386. doi:10.1101/gr.5969107
34. Segata N, Waldron L, Ballarini A, Narasimhan V, Jousson O, Huttenhower C (2012) Metagenomic microbial community profiling using unique clade-specific marker genes. Nat Methods 9:811–814. doi:10.1038/nmeth.2066
35. Brady A, Salzberg S (2011) PhymmBL expanded: confidence scores, custom databases, parallelization and more. Nat Methods 8:367. doi:10.1038/nmeth0511-367
36. Wood DE, Salzberg SL (2014) Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol 15:R46. doi:10.1186/gb-2014-15-3-r46
37. Freitas TAK, Li PE, Scholz MB, Chain PSG (2015) Accurate read-based metagenome characterization using a hierarchical suite of unique signatures. Nucleic Acids Res 43:e69. doi:10.1093/nar/gkv180
38. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403–410. doi:10.1016/S0022-2836(05)80360-2
39. Suzuki S, Kakuta M, Ishida T, Akiyama Y (2015) Faster sequence homology searches by clustering subsequences. Bioinformatics 31:1183–1190. doi:10.1093/bioinformatics/btu780
40. Suzuki S, Kakuta M, Ishida T, Akiyama Y (2014) GHOSTX: an improved sequence homology search algorithm using a query suffix array and a database suffix array. PLoS ONE 9:e103833. doi:10.1371/journal.pone.0103833
41. Kent WJ (2002) BLAT—the BLAST-like alignment tool. Genome Res 12:656–664. doi:10.1101/gr.229202
42. Finn RD, Clements J, Eddy SR (2011) HMMER web server: interactive sequence similarity searching. Nucleic Acids Res 39:W29–W37. doi:10.1093/nar/gkr367
43. Pruitt KD, Tatusova T, Maglott DR (2005) NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res 33:D501–D504. doi:10.1093/nar/gkr367
44. Kanehisa M, Goto S (2000) KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res 28:27–30. doi:10.1093/nar/27.1.29
45. Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, Pang N, Forslund K, Ceric G, Clements J, Heger A, Holm L, Sonnhammer EL, Eddy SR, Bateman A, Finn RD (2012) The Pfam protein families database. Nucleic Acids Res 40:D290–D301. doi:10.1093/nar/gkt1223
46. Bairoch A, Apweiler R (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic Acids Res 28:45–48. doi:10.1093/nar/28.1.45
47. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL (2004) Versatile and open software for comparing large genomes. Genome Biol 5:R12. doi:10.1186/gb-2004-5-2-r12
48. Smith AD, Xuan Z, Zhang MQ (2008) Using quality scores and longer reads improves accuracy of Solexa read mapping. BMC Bioinformatics 9:128. doi:10.1186/1471-2105-9-128
49. Alkan C, Kidd JM, Marques-Bonet T, Aksay G, Antonacci F, Hormozdiari F, Kitzman JO, Baker C, Malig M, Mutlu O, Sahinalp SC, Gibbs RA, Eichler EE (2009) Personalized copy number and segmental duplication maps using next-generation sequencing. Nat Genet 41:1061–1067. doi:10.1038/ng.437
50. Rosen GL, Reichenberger ER, Rosenfeld AM (2011) NBC: the Naive Bayes classification tool webserver for taxonomic classification of metagenomic reads. Bioinformatics 27:127–129. doi:10.1093/bioinformatics/btq619
51. Alneberg J, Bjarnason BS, de Bruijn I, Schirmer M, Quick J, Ijaz UZ, Lahti L, Loman NJ, Andersson AF, Quince C (2014) Binning metagenomic contigs by coverage and composition. Nat Methods 11:1144–1146. doi:10.1038/nmeth.3103
52. Nielsen HB, Almeida M, Juncker AS, Rasmussen S, Li J, Sunagawa S, Plichta DR, Gautier L, Pedersen AG, Le Chatelier E, Pelletier E, Bonde I, Nielsen T, Manichanh C, Arumugam M, Batto JM, Quintanilha Dos Santos MB, Blom N, Borruel N, Burgdorf KS, Boumezbeur F, Casellas F, Doré J, Dworzynski P, Guarner F, Hansen T, Hildebrand F, Kaas RS, Kennedy S, Kristiansen K, Kultima JR, Léonard P, Levenez F, Lund O, Moumen B, Le Paslier D, Pons N, Pedersen O, Prifti E, Qin J, Raes J, Sørensen S, Tap J, Tims S, Ussery DW, Yamada T, MetaHIT Consortium, Renault P, Sicheritz-Ponten T, Bork P, Wang J, Brunak S, Ehrlich SD, MetaHIT Consortium (2014) Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. Nat Biotechnol 32:822–828. doi:10.1038/nbt.2939
53. Sangwan N, Xia F, Gilbert JA (2016) Recovering complete and draft population genomes from metagenome datasets. Microbiome 4:8. doi:10.1186/s40168-016-0154-5
54. Wu M, Scott AJ (2012) Phylogenomic analysis of bacterial and archaeal sequences with AMPHORA2. Bioinformatics 28:1033–1034. doi:10.1093/bioinformatics/bts079
55. Dupont CL, Rusch DB, Yooseph S, Lombardo M-J, Richter RA, Valas R, Novotny M, Yee-Greenbaum J, Selengut JD, Haft DH, Halpern AL, Lasken RS, Nealson K, Friedman R, Venter JC (2012) Genomic insights to SAR86, an abundant and uncultivated marine bacterial lineage. ISME J 6:1186–1199. doi:10.1038/ismej.2011
56. Nayfach S, Pollard KS (2015) Average genome size estimation improves comparative metagenomics and sheds light on the functional ecology of the human microbiome. Genome Biol 16:51. doi:10.1186/s13059-015-0611-7
57. Beszteri B, Temperton B, Frickenhaus S, Giovannoni SJ (2010) Average genome size: a potential source of bias in comparative metagenomics. ISME J 4:1075–1077
58. Raes J, Korbel JO, Lercher MJ, von Mering C, Bork P (2007) Prediction of effective genome size in metagenomic samples. Genome Biol 8:R10. doi:10.1186/gb-2007-8-1-r10
59. Morris JJ, Lenski RE, Zinser ER (2012) The Black Queen Hypothesis: evolution of dependencies through adaptive gene loss. MBio 3:e00036-12
60. Angly FE, Willner D, Prieto-Davó A, Edwards RA, Schmieder R, Vega-Thurber R, Antonopoulos DA, Barott K, Cottrell MT,

Desnues C, Dinsdale EA, Furlan M, Haynes M, Henn MR, Hu Y, Kirchman DL, McDole T, McPherson JD, Meyer F, Miller RM, Mundt E, Naviaux RK, Rodriguez-Mueller B, Stevens R, Wegley L, Zhang L, Zhu B, Rohwer F (2009) The GAAS metagenomic tool and its estimations of viral and microbial average genome size in four major biomes. PLoS Comput Biol 5:e1000593. doi:10.1371/journal.pcbi.1000593

61. Walter J, Ley R (2011) The human gut microbiome: ecology and recent evolutionary changes. Annu Rev Microbiol 65:411–429. doi:10.1146/annurev-micro-090110-102830
62. Wu D, Jospin G, Eisen JA (2013) Systematic identification of gene families for use as "markers" for phylogenetic and phylogeny-driven ecological studies of bacteria and archaea and their major subgroups. PLoS ONE 8:e77033. doi:10.1371/journal.pone.0077033
63. Xu J, Yanagisawa Y, Tsankov AM, Hart C, Aoki K, Kommajosyula N, Steinmann KE, Bochicchio J, Russ C, Regev A, Rando OJ, Nusbaum C, Niki H, Milos P, Weng Z, Rhind N (2012) Genome-wide identification and characterization of replication origins by deep sequencing. Genome Biol 13:R27. doi:10.1186/gb-2012-13-4-r27
64. Skovgaard O, Bak M, Løbner-Olesen A, Tommerup N (2011) Genome-wide detection of chromosomal rearrangements, indels, and mutations in circular chromosomes by short read sequencing. Genome Res 21:1388–1393. doi:10.1101/gr.117416.110
65. Bremer H, Churchward G (1977) An examination of the Cooper–Helmstetter theory of DNA replication in bacteria and its underlying assumptions. J Theor Biol 69:645–654. doi:10.1016/0022-5193(77)90373-3
66. Gao F, Luo H, Zhang CT (2013) DoriC 5.0: an updated database of oriC regions in both bacterial and archaeal genomes. Nucleic Acids Res 41:D90–D93. doi:10.1093/nar/gks990
67. Human Microbiome Project Consortium (2012) Structure, function and diversity of the healthy human microbiome. Nature 486:207–214. doi:10.1038/nature11234
68. Qin J, Li Y, Cai Z, Li S, Zhu J, Zhang F, Liang S, Zhang W, Guan Y, Shen D, Peng Y, Zhang D, Jie Z, Wu W, Qin Y, Xue W, Li J, Han L, Lu D, Wu P, Dai Y, Sun X, Li Z, Tang A, Zhong S, Li X, Chen W, Xu R, Wang M, Feng Q, Gong M, Yu J, Zhang Y, Zhang M, Hansen T, Sanchez G, Raes J, Falony G, Okuda S, Almeida M, LeChatelier E, Renault P, Pons N, Batto JM, Zhang Z, Chen H, Yang R, Zheng W, Li S, Yang H, Wang J, Ehrlich SD, Nielsen R, Pedersen O, Kristiansen K, Wang J (2012) A metagenome-wide association study of gut microbiota in type 2 diabetes. Nature 490:55–60. doi:10.1038/nature11450
69. Koenig RJ, Peterson CM, Jones RL, Saudek C, Lehrman M, Cerami A (1976) Correlation of glucose regulation and hemoglobin AIc in diabetes mellitus. N Engl J Med 295:417–420. doi:10.1056/NEJM197608192950804
70. Shapiro BJ, Friedman J, Cordero OX, Preheim SP, Timberlake SC, Szabo G, Polz MF, Alm EJ (2012) Population genomics of early events in the ecological differentiation of bacteria. Science 336:48–51. doi:10.1126/science.1218198
71. Denef VJ, Banfield JF (2012) In situ evolutionary rate measurements show ecological success of recently emerged bacterial hybrids. Science 336:462–466. doi:10.1126/science.1218389
72. Cordero OX, Polz MF (2014) Explaining microbial genomic diversity in light of evolutionary ecology. Nat Rev Microbiol 12:263–273. doi:10.1038/nrmicro3218
73. Li S-J, Hua Z-S, Huang L-N, Li J, Shi S-H, Chen L-X, Kuang J-L, Liu J, Hu M, Shu W-S (2014) Microbial communities evolve faster in extreme environments. Sci Rep 4:6205. doi:10.1038/srep06205
74. Ran W, Kristensen DM, Koonin EV (2014) Coupling between protein level selection and codon usage optimization in the evolution of bacteria and archaea. mBio 5:e00956-14. doi:10.1128/mBio.00956-14
75. Vieira-Silva S, Rocha EP (2010) The systemic imprint of growth and its uses in ecological (meta)genomics. PLoS Genet 6:e1000808
76. Rocha EP (2004) Codon usage bias from tRNA's point of view: redundancy, specialization, and efficient decoding for translation optimization. Genome Res 14:2279–2286
77. Sharp PM, Bailes E, Grocock RJ, Peden JF, Sockett RE (2005) Variation in the strength of selected codon usage bias among bacteria. Nucleic Acids Res 33:1141–1153
78. Roller M, Lucić V, Nagy I, Perica T, Vlahoviček K (2013) Environmental shaping of codon usage and functional adaptation across microbial communities. Nucleic Acids Res 41:8842–8852. doi:10.1093/nar/gkt673
79. Karlin S, Mrazek J (2000) Predicted highly expressed genes of diverse prokaryotic genomes. J Bacteriol 182:5238–5250
80. Sharp P, Li W (1987) The codon Adaptation Index—a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res 15:1281–1295
81. Supek F, Vlahovicek K (2005) Comparison of codon usage measures and their applicability in prediction of microbial gene expressivity. BMC Bioinformatics 6:15
82. Perna NT, Plunkett G, Burland V, Mau B, Glasner JD, Rose DJ, Mayhew GF, Evans PS, Gregor J, Kirkpatrick HA, Pósfai G, Hackett J, Klink S, Boutin A, Shao Y, Miller L, Grotbeck EJ, Davis NW, Lim A, Dimalanta ET, Potamousis KD, Apodaca J, Anantharaman TS, Lin J, Yen G, Schwartz DC, Welch RA, Blattner FR (2001) Genome sequence of enterohaemorrhagic *Escherichia coli* O157:H7. Nature 409:529–533. doi:10.1038/35054089
83. Read TD, Peterson SN, Tourasse N, Baillie LW, Paulsen IT, Nelson KE, Tettelin H, Fouts DE, Eisen JA, Gill SR, Holtzapple EK, Okstad OA, Helgason E, Rilstone J, Wu M, Kolonay JF, Beanan MJ, Dodson RJ, Brinkac LM, Gwinn M, DeBoy RT, Madpu R, Daugherty SC, Durkin AS, Haft DH, Nelson WC, Peterson JD, Pop M, Khouri HM, Radune D, Benton JL, Mahamoud Y, Jiang L, Hance IR, Weidman JF, Berry KJ, Plaut RD, Wolf AM, Watkins KL, Nierman WC, Hazen A, Cline R, Redmond C, Thwaite JE, White O, Salzberg SL, Thomason B, Friedlander AM, Koehler TM, Hanna PC, Kolstø AB, Fraser CM (2003) The genome sequence of *Bacillus anthracis* Ames and comparison to closely related bacteria. Nature 423:81–86. doi:10.1038/nature01586
84. Rahman SA, Singh Y, Kohli S, Ahmad J, Ehtesham NZ, Tyagi AK, Hasnain SE (2014) Comparative analyses of nonpathogenic, opportunistic, and totally pathogenic *mycobacteria* reveal genomic and biochemical variabilities and highlight the survival attributes of *Mycobacterium tuberculosis*. mBio 5:e02020-14. doi:10.1128/mBio.02020-14
85. Hu B, Xie G, Lo C-C, Starkenburg SR, Chain PSG (2011) Pathogen comparative genomics in the next-generation sequencing era: genome alignments, pangenomics and metagenomics. Brief Funct Genomics 10:322–333. doi:10.1093/bfgp/elr042
86. Sommer MOA, Dantas G, Church GM (2009) Functional characterization of the antibiotic resistance reservoir in the Human microflora. Science 325:1128–1131. doi:10.1126/science.1176950
87. Kassinen A, Krogius-Kurikka L, Mäkivuokko H, Rinttilä T, Paulin L, Corander J, Malinen E, Apajalahti J, Palva A (2007) The fecal microbiota of irritable bowel syndrome patients differs significantly from that of healthy subjects. Gastroenterology 133:24–33. doi:10.1053/j.gastro.2007.04.005
88. Ley RE, Backhed F, Turnbaugh P, Lozupone CA, Knight RD, Gordon JI (2005) Obesity alters gut microbial ecology. Proc

Natl Acad Sci USA 102:11070–11075. doi:10.1073/pnas.0504978102
89. Salipante SJ, Sengupta DJ, Rosenthal C, Costa G, Spangler J, Sims EH, Jacobs MA, Miller SI, Hoogestraat DR, Cookson BT, McCoy C, Matsen FA, Shendure J, Lee CC, Harkins TT, Hoffman NG (2013) Rapid 16S rRNA next-generation sequencing of polymicrobial clinical samples for diagnosis of complex bacterial infections. PLoS ONE 8:e65226. doi:10.1371/journal.pone.0065226
90. Kuczynski J, Lauber CL, Walters WA, Parfrey LW, Clemente JC, Gevers D, Knight R (2011) Experimental and analytical tools for studying the human microbiome. Nat Rev Genet 13:47–58. doi:10.1038/nrg3129
91. Perez-Losada M, Castro-Nallar E, Bendall ML, Freishtat RJ, Crandall KA (2015) Dual transcriptomic profiling of host and microbiota during health and disease in pediatric asthma. PLoS ONE 10:e0131819. doi:10.1371/journal.pone.0131819
92. Rajendhran J, Gunasekaran P (2011) Microbial phylogeny and diversity: small subunit ribosomal RNA sequence analysis and beyond. Microbiol Res 166:99–110. doi:10.1016/j.micres.2010.02.003
93. Chin C-S, Sorenson J, Harris JB, Robins WP, Charles RC, Jean-Charles RR, Bullard J, Webster DR, Kasarskis A, Peluso P, Paxinos EE, Yamaichi Y, Calderwood SB, Mekalanos JJ, Schadt EE, Waldor MK (2011) The origin of the Haitian cholera outbreak strain. N Engl J Med 364:33–42. doi:10.1056/NEJMoa1012928
94. Gardy JL, Johnston JC, Sui SJH, Cook VJ, Shah L, Brodkin E, Rempel S, Moore R, Zhao Y, Holt R, Varhol R, Birol I, Lem M, Sharma MK, Elwood K, Jones SJM, Brinkman FSL, Brunham RC, Tang P (2011) Whole-genome sequencing and social-network analysis of a tuberculosis outbreak. N Engl J Med 364:730–739. doi:10.1056/NEJMoa1003176
95. Rasko DA, Webster DR, Sahl JW, Bashir A, Boisen N, Scheutz F, Paxinos EE, Sebra R, Chin CS, Iliopoulos D, Klammer A, Peluso P, Lee L, Kislyuk AO, Bullard J, Kasarskis A, Wang S, Eid J, Rank D, Redman JC, Steyert SR, Frimodt-Møller J, Struve C, Petersen AM, Krogfelt KA, Nataro JP, Schadt EE, Waldor MK (2011) Origins of the *E. coli* strain causing an outbreak of hemolytic–uremic syndrome in Germany. N Engl J Med 365:709–717. doi:10.1056/NEJMoa1106920
96. Köser CU, Holden MT, Ellington MJ, Cartwright EJ, Brown NM, Ogilvy-Stuart AL, Hsu LY, Chewapreecha C, Croucher NJ, Harris SR, Sanders M, Enright MC, Dougan G, Bentley SD, Parkhill J, Fraser LJ, Betley JR, Schulz-Trieglaff OB, Smith GP, Peacock SJ (2012) Rapid whole-genome sequencing for investigation of a neonatal MRSA outbreak. N Engl J Med 366:2267–2275. doi:10.1056/NEJMoa1109910
97. Aryee A, Price N (2014) Antimicrobial Stewardship—Can we afford to do without it? Br J Clin Pharmacol 79:173–181. doi:10.1111/bcp.12417
98. Dupont H, Mentec H, Sollet J, Bleichner G (2001) Impact of appropriateness of initial antibiotic therapy on the outcome of ventilator-associated pneumonia. Intensive Care Med 27:355–362. doi:10.1007/s001340000640
99. Didelot X, Bowden R, Wilson DJ, Peto TE, Crook DW (2012) Transforming clinical microbiology with bacterial genome sequencing. Nat Rev Genetics 13:601–612. doi:10.1038/nrg3226
100. Hilton SK, Castro-Nallar E, Pérez-Losada M, Toma I, McCaffrey TA, Hoffman EP, Siegel MO, Simon GL, Johnson WE, Crandall KA (2016) Metataxonomic and metagenomic approaches vs. culture-based techniques for clinical pathology. Front Microbiol 7:484. doi:10.3389/fmicb.2016.00484
101. Whitman WB, Coleman DC, Wiebe WJ (1998) Prokaryotes: the unseen majority. Proc Natl Acad Sci USA 95:6578–6583. doi:10.1073/pnas.95.12.6578
102. Frank DN, St Amand AL, Feldman RA, Boedeker EC, Harpaz N, Pace NR (2007) Molecular-phylogenetic characterization of microbial community imbalances in human inflammatory bowel diseases. Proc Natl Acad Sci USA 104:13780–13785. doi:10.1073/pnas.0706625104
103. Ley RE, Peterson DA, Gordon JI (2006) Ecological and evolutionary forces shaping microbial diversity in the human intestine. Cell 124:837–848. doi:10.1016/j.cell.2006.02.017
104. Stappenbeck TS, Hooper LV, Gordon JI (2002) Developmental regulation of intestinal angiogenesis by indigenous microbes via Paneth cells. Proc Natl Acad Sci USA 99:15451–15455. doi:10.1073/pnas.202604299
105. Hooper LV, Wong MH, Thelin A, Hansson L, Falk PG, Gordon JI (2001) Molecular analysis of commensal host-microbial relationships in the intestine. Science 291:881–884. doi:10.1126/science.291.5505.881
106. Hooper LV, Gordon JI (2001) Commensal host-bacterial relationships in the gut. Science 292:1115–1118. doi:10.1126/science.1058709
107. Hooper LV (2009) Do symbiotic bacteria subvert host immunity? Nat Rev Microbiol 7:367–374. doi:10.1038/nrmicro2114
108. Salzman NH, Underwood MA, Bevins CL (2007) Paneth cells, defensins, and the commensal microbiota: a hypothesis on intimate interplay at the intestinal mucosa. Semin Immunol 19:70–83. doi:10.1016/j.smim.2007.04.002
109. Zhang F, Luo W, Shi Y, Fan Z, Ji G (2012) Should we standardize the 1,700-year-old fecal microbiota transplantation? Am J Gastroenterol 107:1755. doi:10.1038/ajg.2012.251
110. Eiseman B, Silen W, Bascom GS, Kauvar AJ (1958) Fecal enema as an adjunct in the treatment of pseudomembranous enterocolitis. Surgery 44:854–859
111. Lee JW, Lattimer LDN, Stephen S, Borum ML, Doman DB (2015) Fecal microbiota transplantation: a review of emerging indications beyond relapsing clostridium difficile toxin colitis. Gastroenterol Hepatol 11:24–32
112. Wang WL, Xu S-Y, Ren Z-G, Tao L, Jiang J-W, Zheng S-S (2015) Application of metagenomics in the human gut microbiome. World J Gastroenterol 21:803–814. doi:10.3748/wjg.v21.i3.803
113. Wooley JC, Ye Y (2009) Metagenomics: Facts and artifacts, and computational challenges. J Comput Sci Technol 25:71–81. doi:10.1007/s11390-010-9306-4
114. Oulas A, Pavloudi C, Polymenakou P, Pavlopoulos GA, Papanikolaou N, Kotoulas G, Arvanitidis C, Iliopoulos I (2015) Metagenomics: tools and insights for analyzing next-generation sequencing data derived from biodiversity studies. Bioinform Biol Insights 9:75–88. doi:10.4137/BBI.S12462
115. Chevreux B, Pfisterer T, Drescher B, Driesel AJ, Müller WE, Wetter T, Suhai S (2004) Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs. Genome Res 14:1147–1159. doi:10.1101/gr.1917404
116. Treangen TJ, Koren S, Sommer DD, Liu B, Astrovskaya I, Ondov B, Darling AE, Phillippy AM, Pop M (2013) MetAMOS: a modular and open source metagenomic assembly and analysis pipeline. Genome Biol 14:R2. doi:10.1186/gb-2013-14-1-r2
117. Paszkiewicz K, Studholme DJ (2010) De novo assembly of short sequence reads. Brief Bioinform 11:457–472. doi:10.1093/bib/bbq020
118. Zerbino DR, Birney E (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res 18:821–829. doi:10.1101/gr.074492.107
119. Namiki T, Hachiya T, Tanaka H, Sakakibara Y (2012) MetaVelvet: an extension of Velvet assembler to de novo metagenome assembly from short sequence reads. Nucleic Acids Res 40:e155. doi:10.1093/nar/gks678

120. Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJ, Birol I (2009) ABySS: a parallel assembler for short read sequence data. Genome Res 19:1117–1123. doi:10.1101/gr.089532
121. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA (2012) SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol 19:455–477. doi:10.1089/cmb.2012.0021
122. Boisvert S, Raymond F, Godzaridis E, Laviolette F, Corbeil J (2012) Ray Meta: scalable de novometagenome assembly and profiling. Genome Biol 13:R122. doi:10.1186/gb-2012-13-12-r122
123. Peng Y, Leung HC, Yiu SM, Chin FY (2011) Meta-IDBA: a de Novo assembler for metagenomic data. Bioinformatics 27:i94–i101. doi:10.1093/bioinformatics/btr216
124. Howe A, Chain PSG (2015) Challenges and opportunities in understanding microbial communities with metagenome assembly (accompanied by IPython Notebook tutorial). Front Microbiol 6:678. doi:10.3389/fmicb.2015.00678
125. Mende DR, Waller AS, Sunagawa S, Järvelin AI, Chan MM, Arumugam M, Raes J, Bork P (2012) Assessment of metagenomic assembly using simulated next generation sequencing data. PLoS ONE 7:e31386. doi:10.1371/journal.pone.0031386
126. Vázquez-Castellanos JF, García-López R, Pérez-Brocal V, Pignatelli M, Moya A (2014) Comparison of different assembly and annotation tools on analysis of simulated viral metagenomic communities in the gut. BMC Genom 15:37. doi:10.1186/1471-2164-15-37
127. Hunt M, Kikuchi T, Sanders M, Newbold C, Berriman M, Otto TD (2013) REAPR: a universal tool for genome assembly evaluation. Genome Biol 14:R47. doi:10.1186/gb-2013-14-5-r47
128. Qin Y, Yalamanchili HK, Qin J, Yan B, Wang J (2015) The current status and challenges in computational analysis of genomic big data. Big Data Res 2:12–18. doi:10.1016/j.bdr.2015.02.005