5

The evolving microbial taxonomy

5.1 Approaching a closed reference set for routine applications

Direct assignment of metagenomic sequencing reads or assembled genes to an existing set of reference sequences would be a quick way of analyzing the data that could serve clinical needs. 16S rRNA gene amplicon sequencing is currently not a closed reference approach (Fig. 5.1) [2,3]; sequences are clustered into operational taxonomic units (OTUs), and an inferred unique "seed sequence" was then used to map to families, genera or species that are already in the database, leaving a varying portion of unannotable OTUs. Metagenomic shotgun sequencing, which can be completely free of PCR amplification biases [4], further includes eukaryotes and viruses and could map to any part of the microbial genomes. According to Dr. Junjie Qin, when he asked Mr. Shenghui Li in the early 2010s to show taxonomic information for each gene that clustered together according to covariations in abundance among hundreds of samples, it turned out that the genes in the same cluster belonged to the same bacterial species [5]. The underlying physical linkage in the microbial genome that was captured by the covariations should be at the strain level, but half of the clusters were unknown species already, and there was much to be improved both computationally and experimentally.

Nowadays, contigs assembled from a single metagenomic sample could be binned according to sequence composition such as tetranucleotide frequency, but covariations in multiple samples could refine the contigs' coverage information and add to the number of assembled genomes in medium to high quality. The metagenomic assembly algorithms are confused by similar sequences from related microbes [6,7], both in the assembly stage (e.g., sequences of two strains in the same sample) and in the binning of contigs into genomes. Reference genome dataset for the human fecal microbiome was constructed from both cultured isolates and metagenome-assembled genomes (MAGs) [8], then the genomes will be dereplicated on species (or strain) level for genome-resolved analyses (Fig. 5.1).

Investigating Human Diseases with the Microbiome: Metagenomics Bench to Bedside.
https://doi.org/10.1016/B978-0-323-91369-0.00004-2

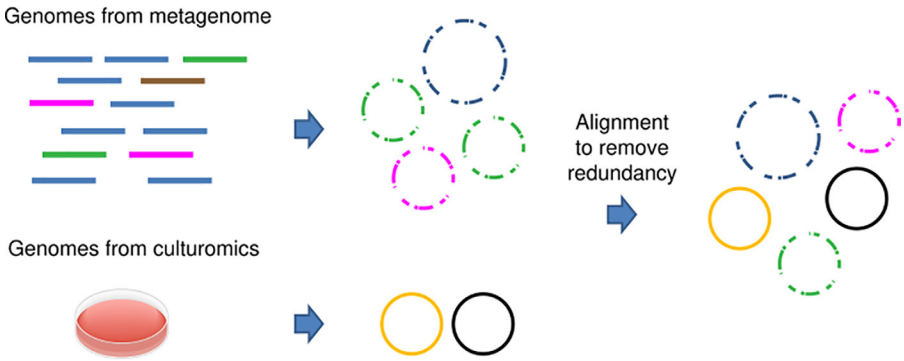

Fig. 5.1 Unbiased retrieval of microbial genomes in metagenomic samples to approach the complete representation of the community. Genomes assembled from high-throughput metagenomic shotgun sequencing data are not yet perfect [1], and are shown in *dashed lines*. Credit: Huijue Jia, Jie Zhu.

The human gut microbiome at the phylum level mostly consists of Firmicutes, Bacteroidetes, Proteobacteria, Actinobacteria, and Fusobacteria, plus Fibrobacteres, Spirochaetes, Lentisphaera, and more mysterious phyla such as Deinococcus-Thermus, Cyanobacteria, Chloroflexi (Fig. 5.2, e.g., db.cngb.org/microbiome/). For the oral microbiome, we see more phyla from the Candidas Phyla Radiation (CPR) (Figs. 5.2 and 5.3), the most famous being the TM7x genus (Saccharibacteria phylum, formerly TM7) which is at least an episymbiont for *Actinomyces odontolyticus* [11–13]. A recent study supports the idea that CPRs evolved through genome reduction from a free-living form, instead of being at the basal place shown in Fig. 5.2 [14], consistent with their obligate symbiont lifestyle and the size constraints [15,16]. Besides the methanogens (Chapter 2, Box 2.4), the human skin contains archaea called *Thaumarchaeota* that can oxidize ammonia to produce nitrite [17].

Due to the traditional view that much of the microbiome is "unculturable" and the difficulty in obtaining the optimal condition, some of the abundant genera and species in the human microbiome got more than 1 draft genomes only recently [18–20]. Comprehensive culturomics for the lungs, for example, would need to take into account the temperature, gas, and pH gradients at the different positions (Chapter 3, Fig. 3.6). The match between culturomics and high-throughput sequencing may have an even longer way to go for fungi (Fig. 5.4), not to mention viruses. While metagenomic sequencing should be able to detect everything in a sample (Chapter 1, Fig. 1.2), it is possible that some low-abundant taxa can be better picked up by specific culturing, which could be facilitated by genomic and other omics information [10,18,21,22].

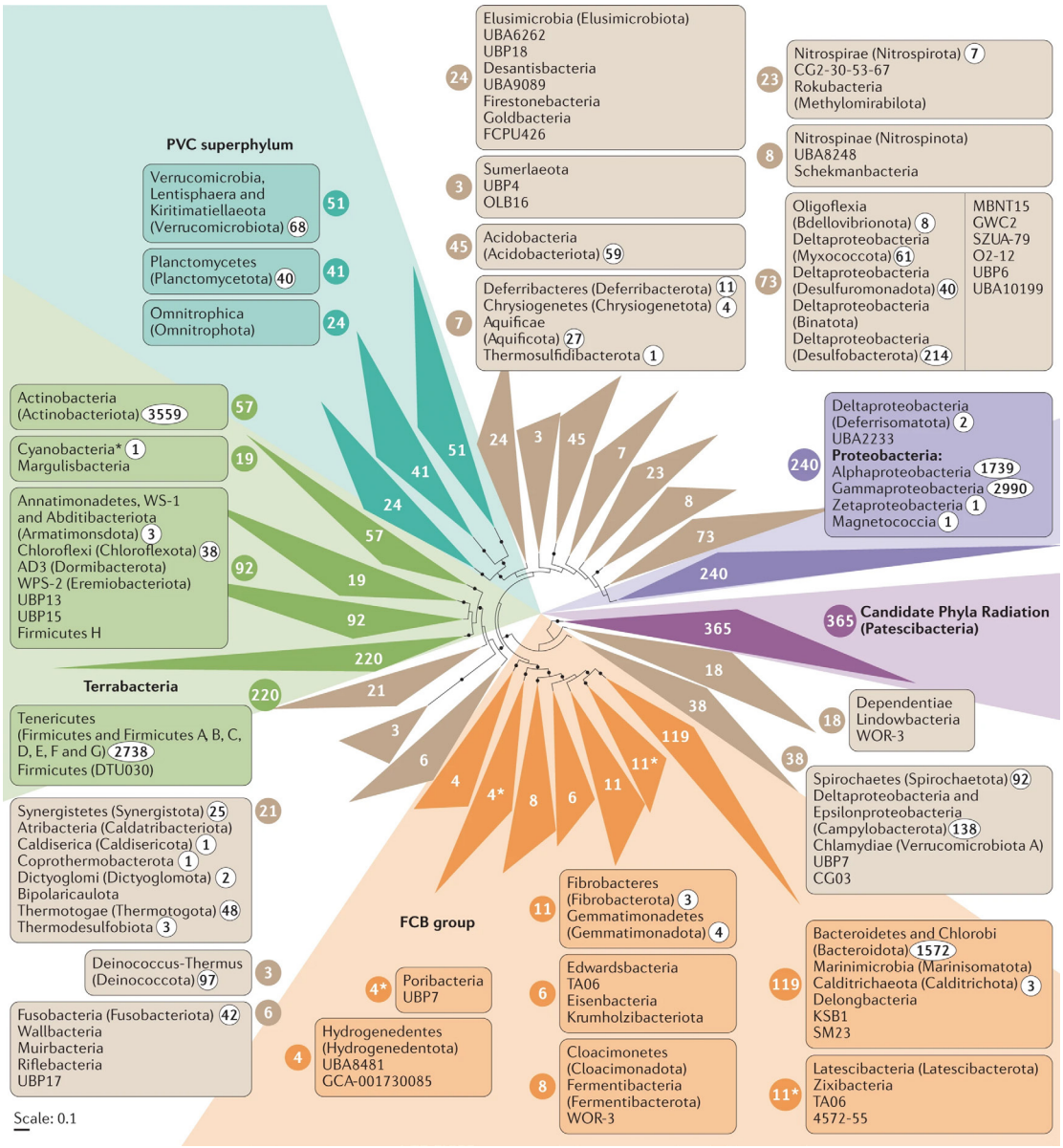

Fig. 5.2 Number of isolated genomes for each bacterial phylum. Without distinguishing between environments, hosts, and body sites, cultured bacteria are currently biased toward Bacteroidetes, Proteobacteria, Firmicutes, and Actinobacteria. A phylogenetic species tree for bacteria, inferred from concatenated alignments of a minimum of 5 out of a total 15 ribosomal proteins per species, encoded by 1541 bacterial genomes that were obtained from the Genome Taxonomy Database [9]. Numbers in *white* font in colored circles are the number of individual taxa in each collapsed clade and are also used to connect corresponding taxa names to clades. Numbers in *black* font in *white*

(Continued)

Fig. 5.2, cont'd ellipses next to taxa names indicate the total number of species level cultured isolates described for those taxa, based on the number of species type strains assigned to each clade that are present in the BacDive database [10] (last accessed 6 April 2020). Taxa without numbers have no cultured isolates recorded in BacDive. Numerous cultured representatives have been reported in the scientific literature that is not represented in the numbers in this figure, because cultures have not been officially described and/or deposited in culture collections, and are therefore not included in BacDive [10] (a comprehensive database recording all cultured bacteria including those not officially described or deposited in culture collections is currently lacking). The tree was generated from datasets containing homologous proteins from the different species included, which were aligned separately using MAFFT (L-INS-i) and the alignments for each protein were then concatenated, such that those proteins belonging to the same species were combined to form a single sequence. Poorly conserved sites in the concatenated alignment were removed using trimAl with the option—gt 0.5. A phylogeny was generated from this trimmed alignment using the model LG + C60 + F + R10 in IQ-TREE with 1000 ultrafast bootstrap replicates. Branches labeled with black dots have support values ≥ 95%. Given the limited protein data set used to infer this phylogeny, in some cases, the deeper relationships between some species or groups may not reflect more widely accepted relationships based on more in-depth and better-supported analyses. Particularly, Deinococcus-Thermus (Deinococcota) and Chlamydiae (Verrucomicrobiota A) do not group with other lineages of Terrabacteria and the PVC superphylum, respectively. *Although numerous cultured representatives for numerous cyanobacterial lineages exist, they are particularly underrepresented in BacDive. Unlike most bacteria, and owing to historical reasons, Cyanobacteria are mostly classified using the Botanical code (i.e., International Code of Nomenclature for algae, fungi, and plants). As a result, Cyanobacteria lack defined type strains and are therefore not extensively listed in BacDive, and a comprehensive database of existing Cyanobacteria cultures is lacking. Credit: From Fig.1 of Lewis WH, Tahon G, Geesink P, Sousa DZ, Ettema TJG. Innovations to culturing the uncultured microbial majority. Nat Rev Microbiol 2021;19:225–40. https://doi.org/10.1038/s41579-020-00458-8.

The commonly used MetaPhlAn series of taxonomic profiling software is based on predetermined marker genes for each taxonomic level, from 50 phyla all the way down to 7677 species and more strains [23–25]. To establish the set of markers, MetaPhlAn2 included 300 archaea genomes, 12,926 bacteria genomes, 3565 virus genomes, and 112 eukaryote genomes, and the total number increased to 99.2 k high-quality genomes in MetaPhlAn3, which has a new set of marker genes and estimates the proportion of unknown taxa [23,26].

Hopefully, family and genus information would soon be more accurate for most metagenomic studies [6,7], and updates on the reference databases and bioinformatics pipelines would not need to be too frequent in the future. As mentioned in Chapter 1, even the Neanderthal oral and fecal microbiome showed many of the same genera we have, and we would love to try to assemble the microbial genomes for such rare samples [27–30], however fragmented. For specific applications (Chapters 7 and 8), a smaller reference database could mean faster and less confusing results.

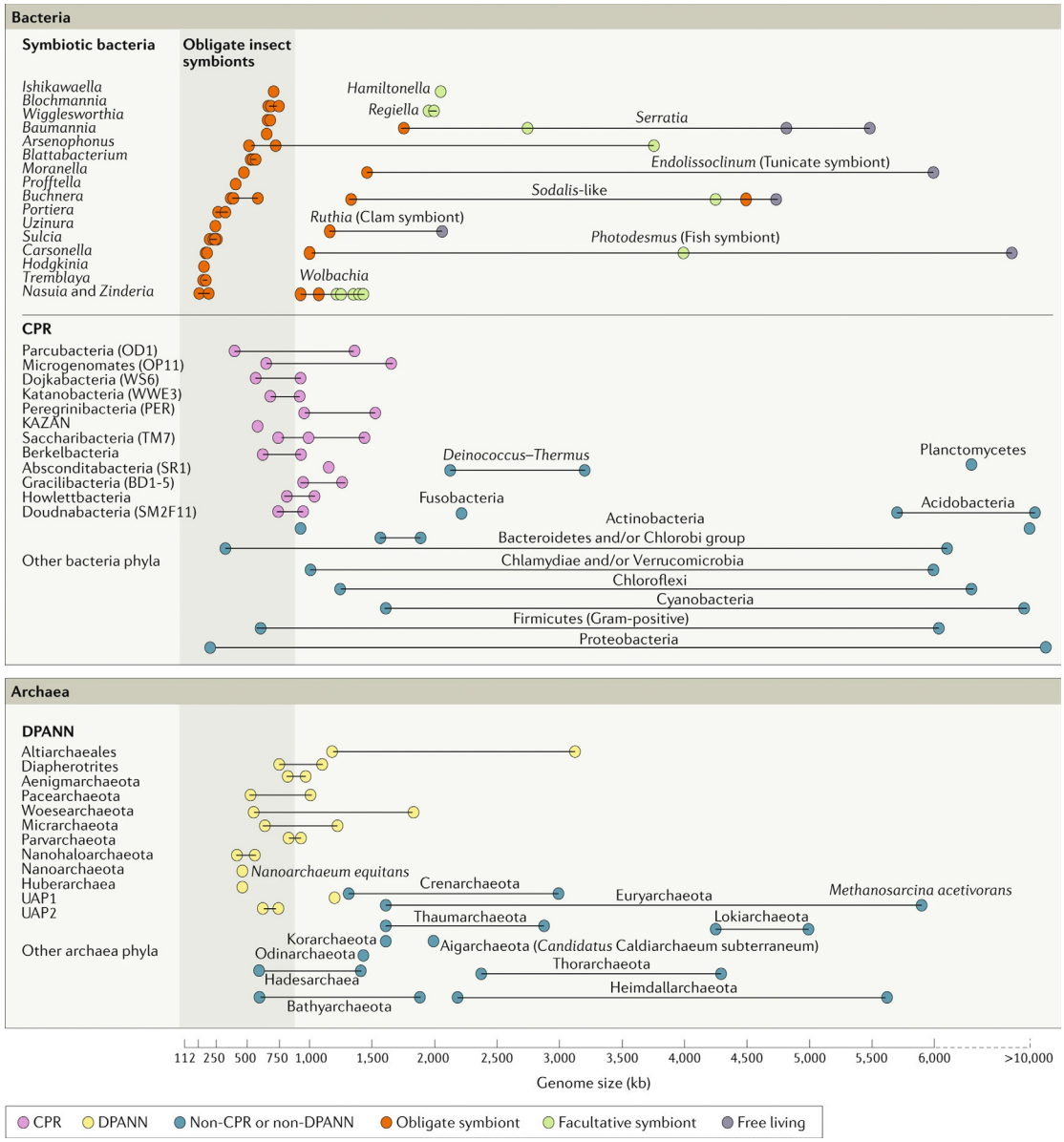

Fig. 5.3 The size ranges for CPR (candidate phyla radiation) bacteria and DPANN ("Candidatus Diapherotrites," "Candidatus Parvarchaeota," "Candidatus Aenigmarchaeota," Nanoarchaeota, "Candidatus Nanohaloarchaeota," and other lineages) archaea genomes compared with size ranges for the genomes of known bacterial symbionts as well as other bacteria and archaea. The top panel shows data for well-studied bacteria that are obligate symbionts *(orange dots)*, facultative symbionts *(green dots)*, and free-living *(gray dots)*. The middle panel provides information for CPR *(purple dots)*, and the bottom panel provides genome size information for DPANN *(yellow dots)*. The middle and bottom panels also show the size ranges for other bacteria and archaea *(blue dots)*. CPR and DPANN genome sizes overlap with those of obligate symbionts. Credit: Fig. 2 of Castelle CJ, Brown CT, Anantharaman K, Probst AJ, Huang RH, Banfield JF. Biosynthetic capacity, metabolic variety and unusual biology in the CPR and DPANN radiations. Nat Rev Microbiol 2018;16:629–45. https://doi.org/10.1038/s41579-018-0076-2.

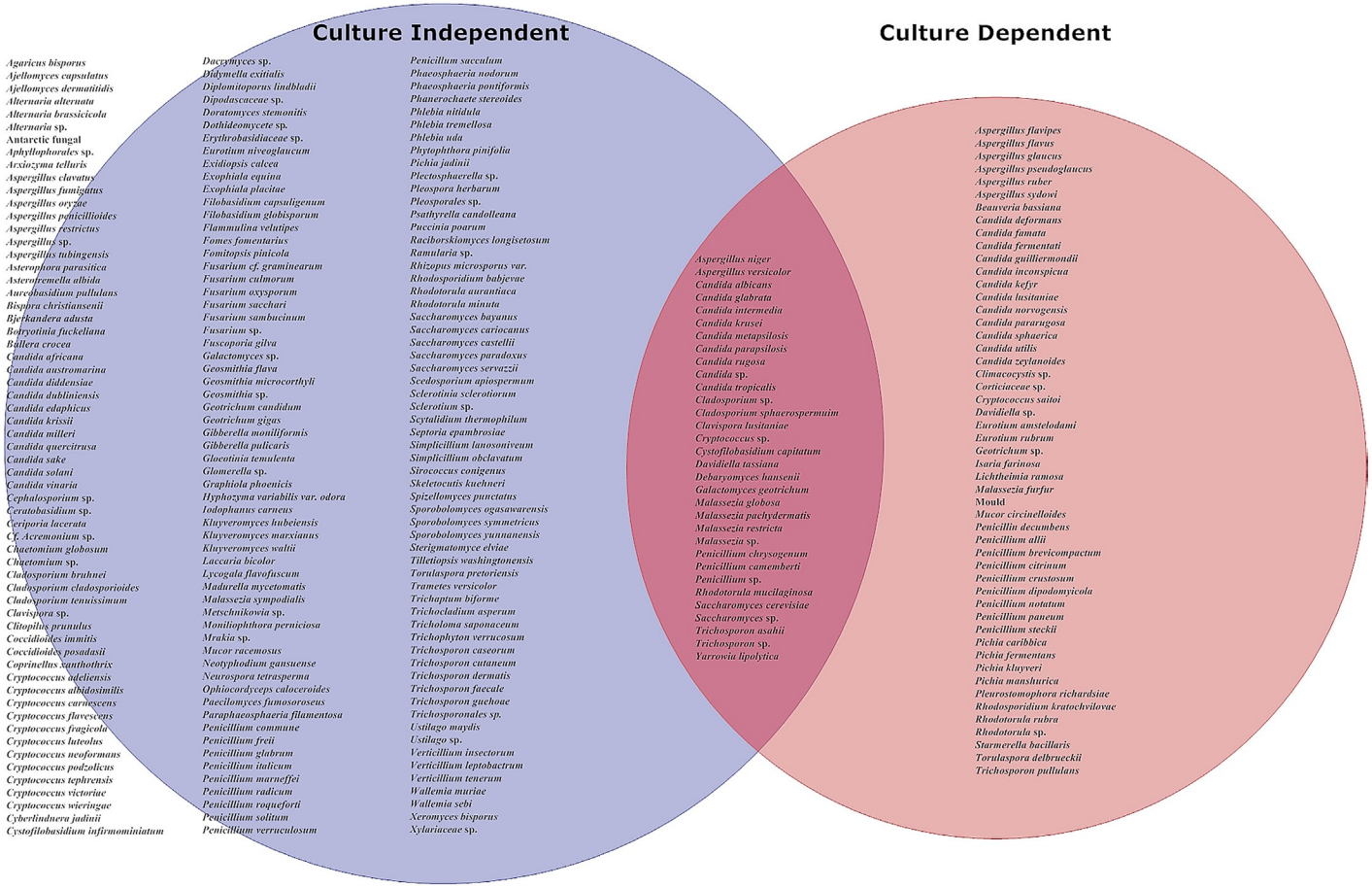

Fig. 5.4 Commonalities and differences between fungal data (at the species level) reported in gut mycobiome studies using culture-dependent and culture-independent methodologies. This Venn diagram highlights fungal species detected by culture-independent only, culture-dependent only, and species that have been detected by both methods (intersection). Credit: Fig. 2 of Huseyin CE, O'Toole PW, Cotter PD, Scanlan PD. Forgotten fungi—the gut mycobiome in human health and disease. FEMS Microbiol Rev 2017;41:479–511. https://doi.org/10.1093/femsre/fuw047.

Worked sample 5.1

Below are the relative abundances of microbial species in one nasal sample (Ju et al. unpublished) using existing softwares, including the k-mer (strings of k nucleotides in the genome)-based Kraken2 + Bracken [31], and the marker gene-based mOTU2 [24], MetaPhlAn2 and MetaPhlAn3. Please get a sense of the different methods and databases, their sensitivity, and accuracy [6,7]. Would you prefer genus-level results instead?

For each method, try to rank the most abundant species 1, 2, 3, … , 30, … Do the different methods look more consistent now?

With more samples, do you see why the most often used correlation coefficient for metagenomic data (between taxa, which would be compositional, i.e., constrained by a total of 1 (Chapter 3), and between a taxon and another phenotype or a taxon at a different body site) is rank-based, such as Spearman's rho (e.g., Chapter 2, Fig. 2.14, the *Bacteroides* vs *Prevotella* is preserved after adjustment for compositional effect [32]), instead of the faster and not so appropriate Pearson's correlation (including SparCC which takes into account the compositional nature of relative abundance data [33])?

What would be the taxa you care more about for your study of the human microbiome, and do you think specific improvements would be needed? Can you guess which method would more easily accommodate newly assembled genomes?

Broad category	Species	Kraken2 + Bracken	mOTU2	MetaPhlAn2	MetaPhlAn3
Bacteria	*Acinetobacter baumannii*	0.157951	0	0	0
Bacteria	*Anaerococcus* species incertae sedis (uncertain placement of species) [meta mOTU v25 12712]	0	0.454865	0	0
Bacteria	Bacilli sp. [ref mOTU v25 00344]	0	0.239723	0	0
Bacteria	*Bacillus cereus*	0.130481	0	0	0
Bacteria	Bacteria sp. [ref mOTU v25 00259]	0	0.155647	0	0
Bacteria	Bacteria sp. [ref mOTU v25 00964]	0	0.131472	0	0
Bacteria	*Brachybacterium paraconglomeratum*	0	0.116516	0	0
Bacteria	*Corynebacterium accolens*	0	24.27697	9.31315	32.04989
Bacteria	*Corynebacterium ammoniagenes*	0.1528	0	0	0
Bacteria	*Corynebacterium aurimucosum*	0.448099	0.490033	0	0
Bacteria	*Corynebacterium camporealensis*	0.458401	0	0	0
Bacteria	*Corynebacterium casei*	0.357106	0	0	0
Bacteria	*Corynebacterium diphtheriae*	0.559695	0	0	0
Bacteria	*Corynebacterium flavescens*	0.375992	0	0	0
Bacteria	*Corynebacterium glutamicum*	0.923669	0	0	0
Bacteria	*Corynebacterium jeikeium*	0.243794	0	0	0

Continued

Broad category	Species	Kraken2 + Bracken	mOTU2	MetaPhlAn2	MetaPhlAn3
Bacteria	*Corynebacterium kroppenstedtii*	1.222402	1.712234	2.09802	1.23582
Bacteria	*Corynebacterium minutissimum*	0.293582	0	0	0
Bacteria	*Corynebacterium phocae*	0.1528	0	0	0
Bacteria	*Corynebacterium propinquum*	0	0.421024	2.12244	0
Bacteria	*Corynebacterium pseudogenitalium*	0	0	0.71378	0
Bacteria	*Corynebacterium resistens*	0.104728	0	0	0
Bacteria	*Corynebacterium simulans*	0.882464	0	0	0
Bacteria	*Corynebacterium singulare*	0.336504	0	0	0
Bacteria	*Corynebacterium* sp. [ref mOTU v25 03067]	0	1.811982	0	0
Bacteria	*Corynebacterium* sp. [ref mOTU v25 00802]	0	0.109947	0	0
Bacteria	*Corynebacterium stationis*	0.14765	0	0	0
Bacteria	*Corynebacterium striatum*	1.857638	0	0	0
Bacteria	*Corynebacterium ureicelerivorans*	0.108162	0	0	0
Bacteria	*Cutibacterium* (formerly *Propionibacterium*) *acnes*	12.71332	9.133923	20.45757	11.96932
Viruses	*Propionibacterium* phage BruceLethal	0.157951	0	0	0
Viruses	*Propionibacterium* phage Moyashi	0.243794	0	0	0
Viruses	*Propionibacterium* phage P101A	0	0	5.99864	0
Viruses	*Propionibacterium* phage PA1-14	0.140782	0	0	0
Viruses	*Propionibacterium* phage PHL009	0.255812	0	0	0
Viruses	*Propionibacterium* phage PHL010M04	0.456684	0	0	0
Viruses	*Propionibacterium* phage PHL030	0.441232	0	0	0
Viruses	*Propionibacterium* phage PHL055	0.104728	0	0	0
Viruses	*Propionibacterium* phage PHL070	0.489304	0	0	0
Viruses	*Propionibacterium* phage PHL082	0.118463	0	0	0
Viruses	*Propionibacterium* phage PHL085	0.396594	0	0	0
Viruses	*Propionibacterium* phage PHL116	0.400028	0	0	0
Viruses	*Propionibacterium* phage PHL132	0.209456	0	0	0
Viruses	*Propionibacterium* phage PHL141	0.157951	0	0	0
Viruses	*Propionibacterium* phage PHL152	0.679875	0	0	0
Viruses	*Propionibacterium* phage PHL171	0.127047	0	0	0
Viruses	*Propionibacterium* phage QueenBey	0.427497	0	0	0
Viruses	*Propionibacterium* virus Attacne	0.260962	0	0	0
Viruses	*Propionibacterium* virus Lauchelly	0.496171	0	0	0
Viruses	*Propionibacterium* virus Ouroboros	0.454967	0	0	0
Viruses	*Propionibacterium* virus P100A	0.108162	0	0	0
Viruses	*Propionibacterium* virus PHL071N05	0.199155	0	0	0
Viruses	*Propionibacterium* virus PHL114L00	0.204306	0	0	0
Viruses	*Propionibacterium* virus Pirate	0.338221	0	0	0
Viruses	*Propionibacterium* virus Solid	0.116746	0	0	0

Continued

Broad category	Species	Kraken2 + Bracken	mOTU2	MetaPhlAn2	MetaPhlAn3
Viruses	*Propionibacterium* virus Stormborn	0.257528	0	0	0
Bacteria	*Cutibacterium granulosum*	0.990626	0.557325	0	1.05931
Bacteria	*Erythrobacteraceae* bacterium CCH12-C2	0	0.11621	0	0
Bacteria	*Haemophilus parainfluenzae*	0.111596	0	0.16822	0
Bacteria	*Klebsiella michiganensis/oxytoca*	0	0.124238	0	0
Bacteria	*Klebsiella oxytoca*	0.121897	0	0	0
Bacteria	*Lautropia mirabilis*	0	0.118127	0.16232	0
Bacteria	*Lawsonella clevelandensis*	0.175119	6.701121	0	0
Eukaryota-Fungi	*Malassezia restricta*	0	0	0	15.47571
Eukaryota-Fungi	*Malassezia* species incertae sedis [meta mOTU v25 12989]	0	15.03185	0	0
Bacteria	*Moraxellaceae* sp. [ref mOTU v25 06002]	0	0.109049	0	0
Bacteria	*Morococcus cerebrosus*	0	0.105722	0	0
Bacteria	*Neisseria elongata*	0.582014	0.283177	0.38064	0.42907
Bacteria	*Neisseria macacae*	0	0	0.1394	0.11948
Bacteria	*Neisseria meningitidis*	0.103011	0.231828	0	0
Bacteria	*Neisseria mucosa*	0.255812	0	0	0
Bacteria	*Neisseria sicca*	0.281564	0	0.43785	0.89061
Bacteria	*Neisseria sicca/macacae*	0	0.216081	0	0
Bacteria	*Neisseria* sp. [ref mOTU v25 04798]	0	0.169086	0	0
Bacteria	*Neisseria* sp. HMSC064E01	0	0.295144	0	0
Bacteria	*Neisseria* sp. oral taxon 014	0	0.102625	0	0
Bacteria	*Neisseria* unclassified	0	0	0.60812	0
Bacteria	*Prevotella melaninogenica*	0	0.196181	0	0
Bacteria	*Pseudomonas stutzeri*	0.18027	0.133901	0	0
Bacteria	Sphingomonadales bacterium RIFCSPHIGHO2 01 FULL 65 20	0	0.101038	0	0
Bacteria	*Staphylococcus aureus*	0.479003	0	0	0
Bacteria	*Staphylococcus capitis*	0.20774	0	0	0
Bacteria	*Staphylococcus epidermidis*	39.33489	33.71282	55.81221	35.78187
Bacteria	*Staphylococcus hominis*	0.307317	0	0.22698	0.16856
Viruses	*Staphylococcus* phage StB27	0.259245	0	0	0
Viruses	*Staphylococcus* virus IPLAC1C	0.382859	0	0	0
Viruses	*Staphylococcus* virus SEP9	7.926725	0	0	0
Viruses	*Staphylococcus* virus Sextaec	12.46781	0	0	0
Bacteria	*Streptococcus mitis*	0.173403	0	0	0.15237
Bacteria	*Streptococcus mitis/oralis/pneumoniae*	0	0	0.25495	0
Bacteria	*Streptococcus sanguinis/cristatus*	0	0.106522	0	0
Bacteria	*Streptococcus* sp. [ref mOTU v25 00283]	0	0.412096	0	0

mOTU2 was based on 5232 ref mOTUs which have reference genomes, and 2494 meta mOTUs which were supported by metagenomic data from the human or ocean microbiome.

Worked sample 5.2

Take a sample of the tongue dorsum and a sample of the saliva from the same person for metagenomic shotgun sequencing, or find some published data.

How many species and higher taxa do you get in each sample?

Could you assemble some high-quality genomes, and how different are they between the tongue and the saliva samples?

Is the lactic acid-metabolizing *Veillonella* found in the gut of elite runners [34] different from the *Veillonella* in autoimmune diseases such as rheumatoid arthritis (Section 4.4.1)?

5.2 Sparser data with increasing taxonomic resolution

We talked about statistical practices with metagenomic studies in Chapter 3. While knowing the species (Box 5.1) or strains are important for functional characterizations, finer grains of taxonomy means that more samples would have zero abundance for a taxon. This sparsity is a problem in statistics. There are efforts to distinguish sampling zeros (e.g., not detected due to low sequencing amount) from real absence (structural zeros) [35]. Before better statistical methods are developed for metagenomic data, we are going to see some larger *P*-values with a higher-taxonomic resolution, for which the effective sample size is smaller. For example, the phylum Proteobacteria showed up multiple times in an MR (Mendelian Randomization) analysis of fecal microbiome and plasma metabolites, often more "significant" than the genera or species [36]. Below the genus level, the issue of the same sequencing read mapping to multiple genomes can strongly affect the relative abundance values [6]. So we are not sure whether multiples species or strains are similarly associated with a human gene, or there is finer work to do in the abundance profile, before we try it.

The finer taxonomic resolution would also affect correlations (between taxa, or between a taxon and another omics feature) calculated from multiple samples. A taxon that was ranked as highly abundant (e.g., Worked sample 5.1) would be broken into smaller pieces with more fluctuating relative abundances and may no longer show up in Spearman's correlation or other statistical measures.

To select biomarkers in metagenome-wide association studies (MWAS), machine learning algorithms such as random forest and LASSO (Least absolute shrinkage and selection operator) can handle the sparse (zero-inflated) data [37], and the microbial markers selected by the models are more like to be validated in a different cohort, instead of being overfit to the training set. Moreover, 10- or 5-fold cross-validation is often used with random forest models (RFcv) and run multiple times, i.e., 1/10 of the samples were randomly left out in the training, and used as a test set, the second time another 1/10

of the samples were left out for validation, the third time... (e.g., Refs. [38–40]).

Recent algorithm developments in neural networks (Artificial Intelligence) could potentially better handle the different layers of relationships in metagenomic data, along with other omics, facilitating association analyses, biomarker discovery for diagnosis and prognosis, as well as MAG assembly and functional annotations.

Box 5.1 The species concept for bacteria

Defining species was not only a problem for bacteria or viruses. Charles Darwin wrote in the Origin of Species published in 1859 [43,44]:

"To sum up, I believe that species come to be tolerably well-defined objects, and do not at any one period present an inextricable chaos of varying and intermediate links …

… if my theory be true, numberless intermediate varieties, linking most closely all the species of the same group together, must assuredly have existed; but the very process of natural selection constantly tends…to exterminate the parent-forms and the intermediate links.

… it will be seen that I look upon the term species, as one arbitrarily given for the sake of convenience to a set of individuals closely resembling each other, and that it does not essentially differ from the term variety, which is given to less distinct and more fluctuating forms.

In short, we shall have to treat species in the same manner as those naturalists treat genera, who admit that genera are merely artificial combinations made for convenience. This may not be a cheering prospect; but we shall at least be freed from the vain search for the undiscovered and undiscoverable essence of the term species."

Alfred Russel Wallace, natural selection's co-discoverer, (he made other comments earlier) later reached a species definition that was more like an ecotype, but is maintained across generations:

"A species … is a group of living organisms, separated from all other such groups by a set of distinctive character(istic)s, having relations to the environment not identical with those of any other group of organisms, and having the power of continuously reproducing its like"—Wallace [45].

A major difficulty for defining species in prokaryotic organisms is that sexual reproduction cannot be used as a criterion to mark boundaries between species. The microbiome has interestingly be shown to facilitate reproductive isolation, mate discrimination, and hybrid infertility/lethality in insects, thereby contributing to speciation [46].

Also of concern is that the microbial genome may be more plastic. Horizontal gene transfer occurs through mobile elements that can cross taxonomic boundaries, yet does not appear to happen too frequently [16,47]. Sequences of core genes do show a clear boundary at the species level (Fig. 5.5). So, bacteria species have distinct genomic features that would not easily shift into a different species through the accumulation of mutations or through horizontal gene transfer. The core genes also define the metabolic and cell wall traits that are traditionally assayed for in microbiology (e.g., Chapter 1, Fig. 1.10).

Ecotypes, a concept from the niche theory, are not necessarily genetic, and also include heterogeneous gene expression at the single-cell level in a community [48–50]. When not aligned with heritable genomic features, ecotypes are more for functional studies under a variety of conditions and are rather auxiliary for defining microbial species. For species in the human microbiome, functional studies would probably need to include interactions with the host immune system, e.g., antigenic properties can be predicted from the microbial genome if we have first accumulated experimental evidence.

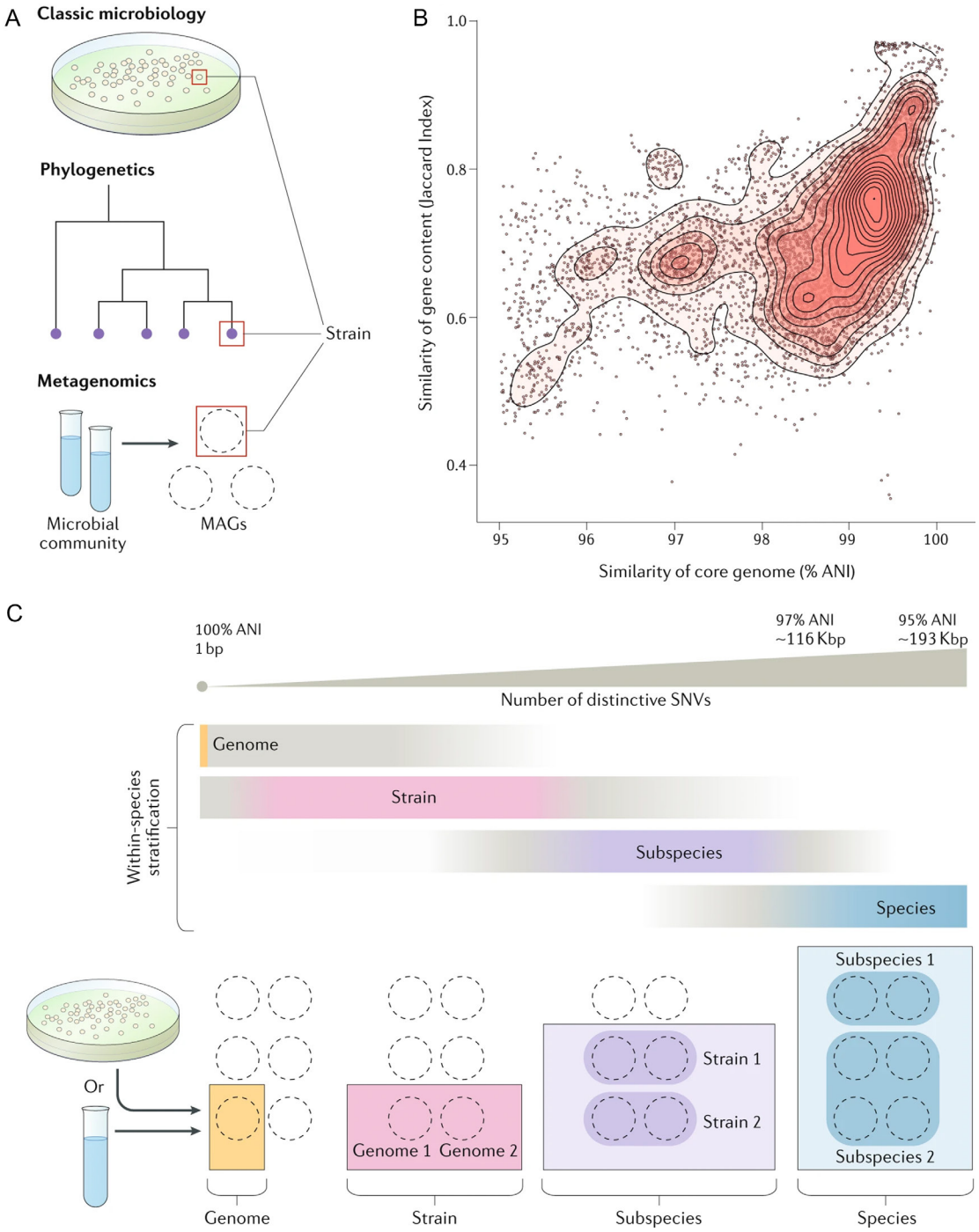

Fig. 5.5 Never the same—Within species diversity of microbial genomes. (A) Different operational definitions of "strain," based on the field of investigation: a cultured isolate in classic microbiology, a leaf node in a phylogenetic tree, and a metagenome-assembled genome (MAG) in metagenomics. (B) Each point is a pairwise comparison of one isolate genome versus all other conspecific isolate genomes. The data are from 155 bacterial species, each with at least 10 sequenced isolate genomes. The opacity of the *red*-colored topographical overlay indicates the density of points. The plot shows the relationship between the similarity of the core genome, measured by average nucleotide identity (ANI), versus the similarity of gene content, measured by the Jaccard Index. Genomes with higher similarity between their core gene sequences tend to have more genes in common (Spearman correlation $R=0.57$, $P<2.2\times10^{-16}$).

(Continued)

Fig. 5.5, cont'd However, a high ANI does not necessarily imply a highly similar gene content, with many genomes with an over 99% core genome ANI having less than 70% of genes in common. Most within-species ANI values are greater than 97%; the few data points below 95% ANI are not shown (83% and 4% of data points, respectively). (C) Spatial distribution of key terminology used to stratify variation within bacterial species, ranging from a single nucleotide variant (SNV) in the whole genome to the species-level threshold (97% ANI). The colored portions of the bars reflect the recommended scope of use for each term, and the *gray* portions indicate the common, often unspecific, scope of use. Broadly speaking, conspecific genomes have identical nucleotides at homologous positions across 97% of their genome (97% ANI), which corresponds to differences in the order of 116,000 SNVs based on an average bacterial genome size of 3.87 Mb. The bottom panel illustrates the hierarchy of these terms, with a species potentially containing multiple subspecies, a subspecies containing multiple strains and a strain containing multiple (nonidentical) genomes. These genomes can be sequenced from cultured isolates or through assembly

Worked sample 5.3

Based on 1267 fecal samples profiled according to a reference gene catalog of 9,879,896 genes (the shorter version of "redundant genes" removed at merging, according to 95% identity), we previously estimated that every two individuals share ~ 1/3 of their gut microbial genes. Each sample contained an average of 762,665 genes and any two samples had in common an average of 250,382 genes (32.8% of 762,665 genes) [41].

At the taxonomic level, and focusing on a particular group of people, what is your current thinking for the number of gut microbial phyla, …, families, genera, and species shared between two individuals? Could you see different patterns in spore-forming bacteria [42], and bacteria that rely more on vertical transmission (e.g., between mother and infant)? (Dispersal limitation, Chapter 2, Fig. 2.3).

What about the microbiome in other body sites?

5.3 Evolutionary history below the species level

Although reproductive isolation does not work for bacteria (Box 5.1), the boundary for species is clear at the genome level based on core genes (Fig. 5.5). Some genera are eventually split from a older genus name and renamed according to genomic distance and functional differences. *Prevotella copri* (Chapter 2), which for years only had a single draft genome, contains at least 4 clades and is now referred to as the *Prevotella copri* complex (Fig. 5.6), to indicate that it is not a homogenous species. For a DNA polymerase error rate of 10^{-8} during genome replication and a repair rate lower than that of eukaryotes, *Escherichia coli* accumulates about 1 mutation in every 1.85×10^{9} nucleotides for a genome of 4.6 Mb [51]. *Bacteroides fragilis* have a genome size of 5.2 Mb, and repeated isolation of the species from the same individuals showed

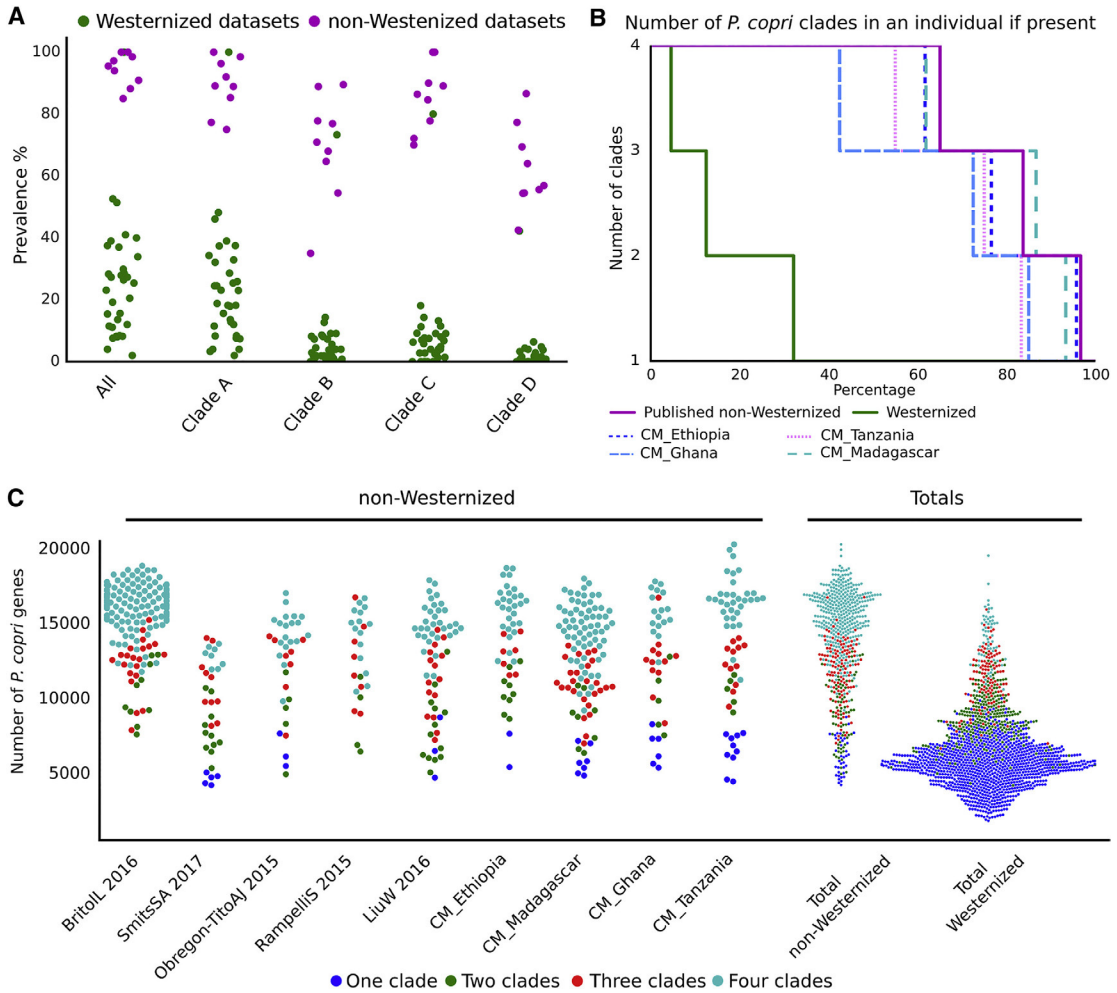

Fig. 5.6 Prevalence of the *Prevotella copri* complex and its association with non-Westernized populations. (A) *Prevotella copri* prevalence in non-Westernized and Westernized datasets. "All" refers to the prevalence of any of the four clades being present. (B) Percentage of individuals harboring multiple *Prevotella copri* clades. (C) *Prevotella copri* complex pangenome sizes for non-Westernized individuals by dataset compared to Westernized individuals. Protein coding genes specific to each clade of the *Prevotella copri* complex were defined as present in > 95% of the *Prevotella copri* genomes of a given clade but absent in all others. This gave for Clade A $n = 430$ markers, for Clade B $n = 954$, for Clade C $n = 479$, and for Clade D $n = 585$. Credit: Fig. 3 of Tett A, Huang KD, Asnicar F, Fehlner-Peach H, Pasolli E, Karcher N, et al. The Prevotella copri complex comprises four distinct clades underrepresented in westernized populations. Cell Host Microbe 2019. https://doi.org/10.1016/j.chom.2019.08.018. of a metagenomic sample, creating a MAG that represents the consensus genome of a population of cells. Credit: Fig. 2 of Van Rossum T, Ferretti P, Maistrenko OM, Bork P. Diversity within species: interpreting strains in microbiomes. Nat Rev Microbiol 2020;18:491–506. https://doi.org/10.1038/s41579-020-0368-1. Panel B was adapted by Van Rossum et al. from Maistrenko OM, Mende DR, Luetge M, Hildebrand F, Schmidt TSB, Li SS, et al. Disentangling the impact of environmental and phylogenetic constraints on prokaryotic within-species diversity. ISME J 2020;14:1247–59. https://doi.org/10.1038/s41396-020-0600-z.

an accumulation of 1 SNP per year [52] (Figs. 5.7 and 5.8), suggesting that this gut mucosal-resident bacterium may only replicate about once every day $(1.85\times10^9/(5.2\times10^6)/365=0.97\approx1)$, if starting from a homogenous population. Greater changes are probably not well tolerated at this niche, e.g., colonization-deficient *Bacteroides fragilis* Δ*ccf* (Chapter 1 Fig. 1.1C), while horizontal gene transfer could occasionally be observed [52]. Those that replicate more often or have a larger population could accumulate SNPs faster. For *Pseudomonas aeruginosa* in the lungs of cystic fibrous patients, distinct lineages could form in different regions of the lungs [53].

Horizontal gene transfers through plasmids, prophages, or other mobile elements are more frequent for functions that provide a clear advantage under stress, e.g., antibiotic resistance [54,55]. Such functions can become a burden when the stress is no longer present, and the abundant strains do not have to be the most resistant ones when the stress is present [56].

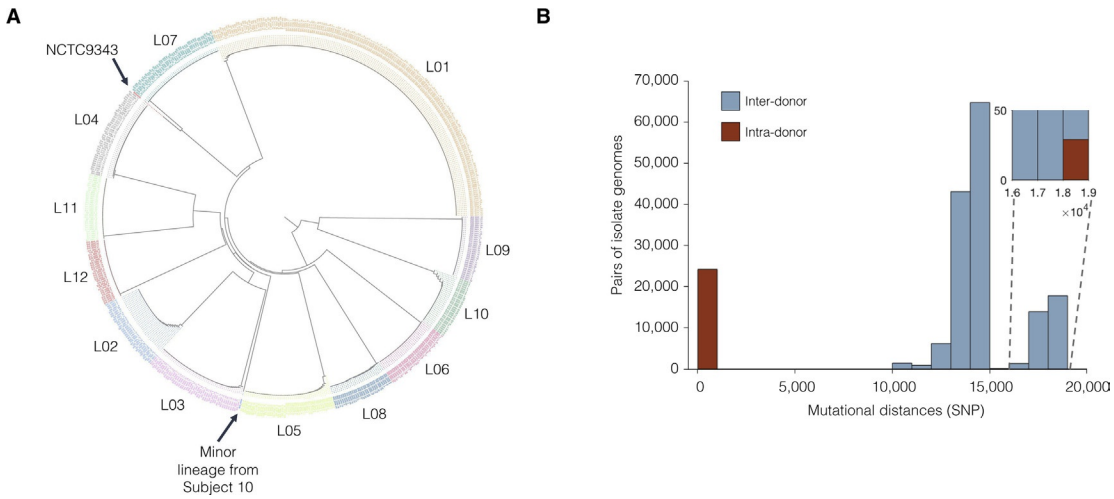

Fig. 5.7 Example of stable lineage and individual specificity of a gut bacterium. Each of the 12 healthy subject's *Bacteroides fragilis* population is dominated by a single lineage. Samples from 7 of the subjects span 2 years. (A) Phylogenetic reconstruction shows that isolates cluster by subject ($n=602$). Isolates are colored according to the subject, which grouped in lineages (L01 to L12). The arrow on top-left marks the NCBI reference genome assembly for *B. fragilis* CCUG4856T (NCTC9343). The arrow next to L05 indicates a single sample from Subject 10 that did not cluster in L10. (B) Isolates from the same subjects generally differ by < 100 single nucleotide differences (SNPs), while isolates from different subjects differ by > 10,000 SNPs. Inset: intra-subject pairs separated by > 18,000 SNPs all involve the outlier isolate from subject 10. Credit: Fig. 1 of Zhao S, Lieberman TD, Poyet M, Kauffman KM, Gibbons SM, Groussin M, et al. Adaptive evolution within gut microbiomes of healthy people. Cell Host Microbe 2019. https://doi.org/10.1016/j.chom.2019.03.007.

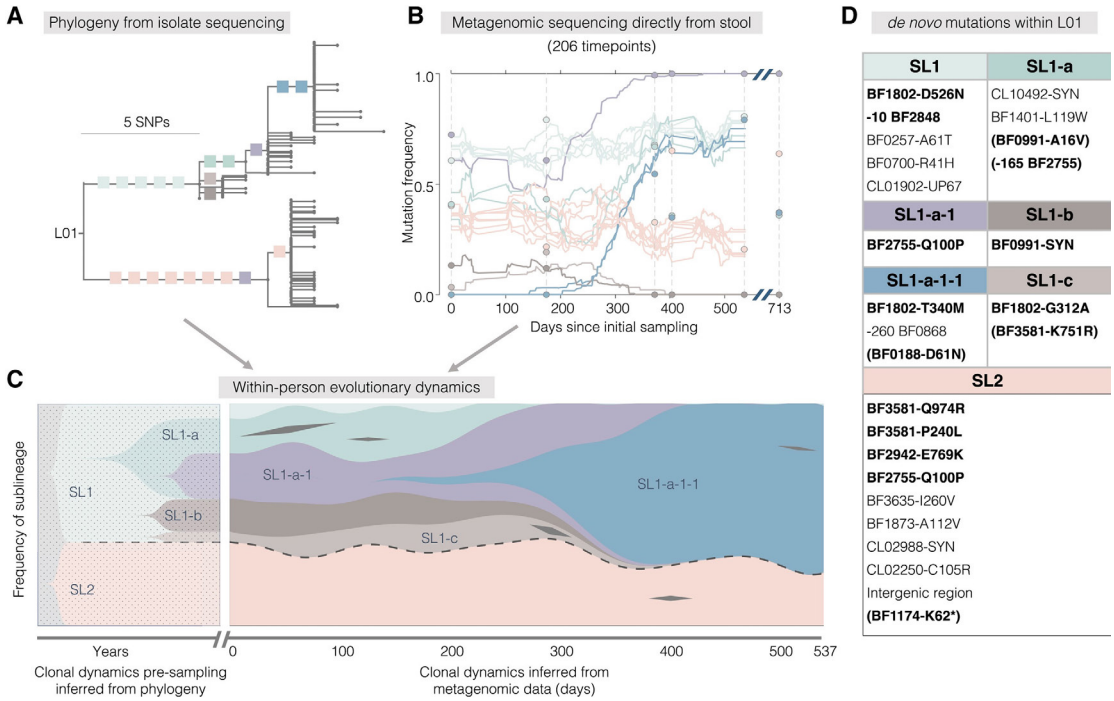

Fig. 5.8 Evolutionary dynamics over a 1.5-year sampling period for one volunteer reveals a steady increase in mutational frequencies and a stable coexistence of two sublineages of *Bacteroides fragilis*. Mobile genetic elements are not shown. (A–C) We combined 206 stool metagenomes and 187 isolate whole genomes to infer evolutionary dynamics within L01 (Fig. 5.7). (A) Branches with at least 4 isolates are labeled with colored *squares* that represent individual SNPs. One SNP was inferred to have happened twice and is indicated in both locations *(purple)*. (B) Frequencies of labeled SNPs were inferred from metagenomes. *Circles* represent SNP frequencies inferred from isolated genomes. (C) We combined these data types to infer the trajectory of sublineages prior to and during sampling. Sublineages are labeled with names and colored as in (A). The two major sublineages, SL1 and SL2, are separated by a *dashed line*. *Black* diamonds represent transient SNPs from polysaccharide utilization (PULs) and cell-envelope biosynthesis genes. (D) The identity of SNPs is shown in (A–C). SNPs in the 16 genes under positive selection are bolded and transient mutations in these genes are indicated with parentheses. Negative numbers indicate mutations upstream of the start of the gene. Credit: Cropped from Fig. 5 of Zhao S, Lieberman TD, Poyet M, Kauffman KM, Gibbons SM, Groussin M, et al. Adaptive evolution within gut microbiomes of healthy people. Cell Host Microbe 2019. https://doi.org/10.1016/j.chom.2019.03.007.

5.4 Whole-cell modeling to predict functional differences from genomic differences?

With a small genome of 525 genes, and over 900 publications, *Mycoplasma genitalium* became the first organism to be computationally modeled for every major process in the cell (Fig. 5.9) [57], with 27.5% of the parameter values found for actual experiments with *M. genitalium* instead of other bacteria. The

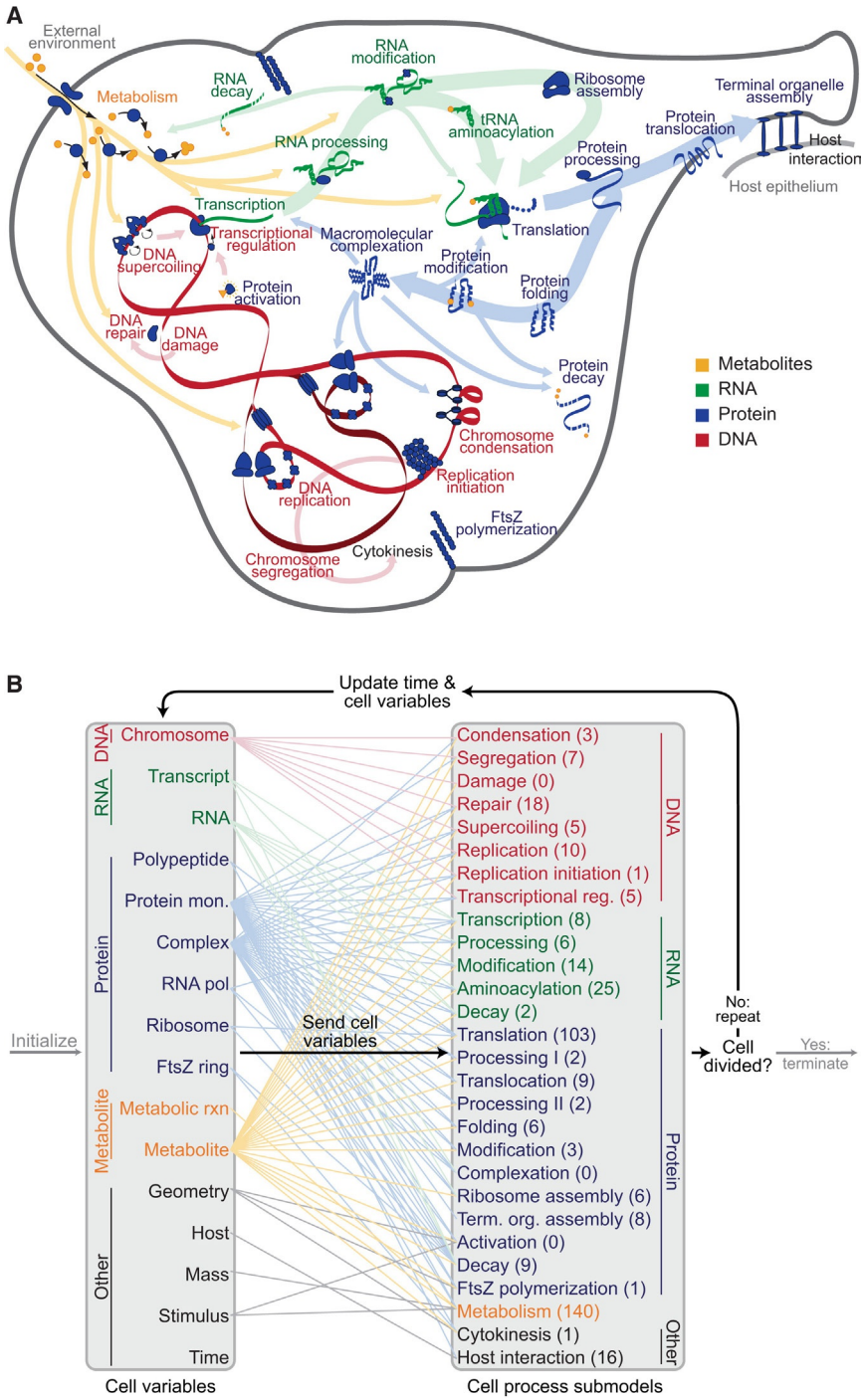

Fig. 5.9 A whole-cell model for *Mycoplasma genitalium*. (A) *M. genitalium* whole-cell model integrates 28 submodels of diverse cellular processes. Diagram schematically depicts the 28 submodels as colored words—grouped by category as metabolic *(orange)*, RNA *(green)*, protein *(blue)*, and DNA *(red)*—in the context of a single *M. genitalium* cell with its characteristic flask-like shape. Submodels are connected through common metabolites, RNA, protein, and the chromosome, which are depicted as *orange, green, blue, and red arrows*, respectively.

(Continued)

Fig. 5.9, cont'd (B) The model integrates cellular function submodels through 16 cell variables. First, simulations are randomly initialized to the beginning of the cell cycle (left *gray arrow*). Next, for each 1 s time step *(dark black arrows)*, the submodels retrieve the current values of the cellular variables, calculate their contributions to the temporal evolution of the cell variables, and update the values of the cellular variables. This is repeated thousands of times during the course of each simulation. For clarity, cell functions and variables are grouped into five physiologic categories: DNA *(red)*, RNA *(green)*, protein *(blue)*, metabolite *(orange)*, and other *(black)*. Colored lines between the variables and submodels indicate the cell variables predicted by each submodel. The number of genes associated with each submodel is indicated in parentheses. Finally, simulations are terminated upon cell division when the septum diameter equals zero (right *gray arrow*). Credit: From Fig.1 of Karr JR, Sanghvi JC, Macklin DN, Gutschow MV., Jacobs JM, Bolival B, et al. A whole-cell computational model predicts phenotype from genotype. Cell 2012;150:389–401. https://doi.org/10.1016/j.cell.2012.05.044.

recent model for *Escherichia coli* considered 1214 genes (43% of the well-annotated *Escherichia coli* genes) and found > 19,000 measured parameter values from decades of literature on *Escherichia coli* itself [58]. This is still a luxury for most members of the human microbiome, which might have only one publication that named the microbe. Such models, being more defined compared to constraint-based models that only consider the metabolic flux [59], could become a key bridge between microbial genomes and phenotypes, and guide further experiments [57,58]. The *M. genitalium* study predicted essential and nonessential genes and identified previously unknown redundant functions from the discrepancy [57]. The *Escherichia coli* model identified discrepancies such as an insufficient number of ribosomes and RNA polymerases to maintain the doubling time, and found many essential proteins to be not transcribed within a cell cycle, and could be absent in some cells [58]. For example, the genes encoding the heterodimeric 4-amino-4-deoxychorismate synthase, *pabA* and *pabB*, are each transcribed with a frequency of 0.94 and 0.66 times per cell cycle, respectively, producing an average of 34 PabA proteins and 101 PabB proteins per generation; in the absence of PabAB heterodimers, cellular 5,10-dimethylene tetrahydrofolate (methylene-THF) decreases over time.

An *Escherichia coli* cell is so densely packed with macromolecules (200–300 g/L [60], higher than the simulations in Fig. 1.3C) that it is in a glass state—molecules larger than ~ 30 nm, such as ribosomes (21 nm each), large enzyme complexes, plasmids, and phage particles, would stay in place with no diffusion, in contrast to freely diffusing small molecules [61] (Fig. 5.10). Metabolic activity fluidizes the glass state and increases movement [61]. The nucleus in mammalian cells

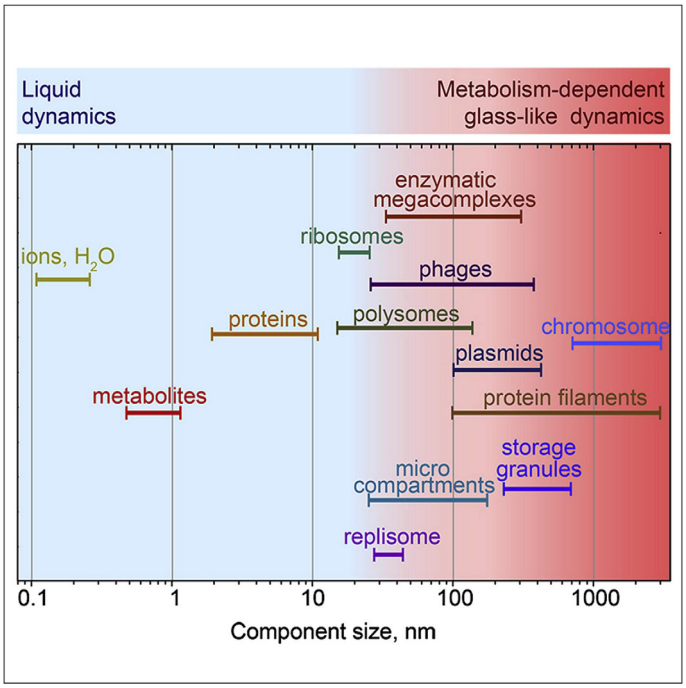

Fig. 5.10 Size of molecules, and mobility in bacterial cells dependent on active metabolism. Credit: Graphical Abstract of Parry BR, Surovtsev IV., Cabeen MT, O'Hern CS, Dufresne ER, Jacobs-Wagner C. The bacterial cytoplasm has glass-like properties and is fluidized by metabolic activity. Cell 2014;156:183–94. https://doi.org/10.1016/j.cell.2013.11.028.

is also much like that. If whole-cell models have enough experimental data to predict physical properties, they would eventually be able to predict dynamics in colonies, and in complex interactions with the host epithelium.

Metabolic models of multiple species can be validated by in vitro experiments and cohort data. *Bacteroides thetaiotaomicron* by itself produces acetate and propionate. In the presence of *Faecalibacterium prausnitzii* or *Eubacterium rectale*, the 4-species community (with *Bifidobacterium adolescentis* and *Ruminococcus bromii*) produces butyrate, acetate, and much less propionate than *Bacteroides thetaiotaomicron* alone [62] (Fig. 5.11, more on the short-chain fatty acids (SCFAs) in Chapter 6 Fig. 6.5). A model of butyrate production considered as many as 25 species and identified inhibition by hydrogen sulfate (H_2S), effect of pH, etc. [63].

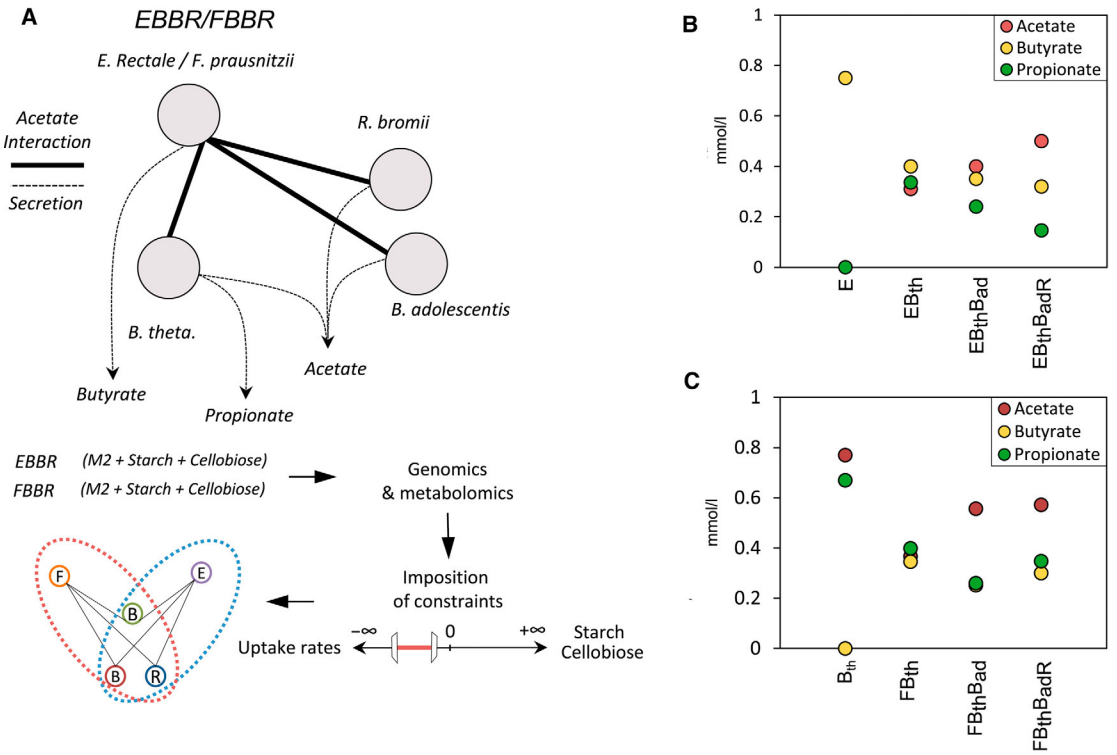

Fig. 5.11 Example of a metabolic model of multiple species. (A) Two in silico microbial communities, EBBR (*Eubacterium rectale* + *Bacteroides thetaiotaomicron* + *Bifidobacterium adolescentis* + *Ruminococcus bromii*) and FBBR (*Faecalibacteirum prausnitzii* + *B. thetaiotaomicron* + *B. adolescentis* + *R. bromii*), were designed and simulated using the CASINO (Community And Systems-level INteractive Optimization) Toolbox. The results were compared with data from triplicate in vitro experiments for EBBR and FBBR communities, grown in M2 medium supplemented with 0.2% (weight/volume) starch and 0.2% (weight/volume) cellobiose. In CASINO, the interactions of the bacteria as well as the phenotype of the community were identified using an optimization algorithm. Growth of each bacterium had local optimum, whereas the community had global optimum. The community optimum was detected by the intersection point of the fixed constraints for the community and the calculated dynamic constraints, which were obtained by summation of the local and community forces. (B and C) Network structure influenced SCFA production. The sensitivity of CASINO optimization was tested by evaluating the changes in the SCFA profile upon adding different species to the community. First, the most important receptor (receiving metabolites from the other microbes) and effector (producing metabolites consumed by receptors) in the communities were identified according to power centrality and degree centrality. 1 mmol/L of glucose was used for all the simulations, and the SCFA profiles were predicted. Following identification of the dominant receptor and effector, the key species, the other species were added to the community one by one until the EBBR (B) and FBBR (C) communities were reconstructed. Comparison between the simulations showed that the SCFA profile is very sensitive to the absence and presence of species with respect to their abundance and interactions. Credit: Fig. 2A, Fig. 3B of Shoaie S, Ghaffari P, Kovatcheva-Datchary P, Mardinoglu A, Sen P, Pujos-Guillot E, et al. Quantifying diet-induced metabolic changes of the human gut microbiome. Cell Metab 2015;22:320–31. https://doi.org/10.1016/j.cmet.2015.07.001.

5.5 Summary

The human microbiome involves the largest taxonomic division, from all domains of life all the way down to microbial strains, and single-nucleotide polymorphisms (SNPs). With reference genomes accumulating from both cultured isolates and metagenomic assembly for each body site, taxonomic profilers would have more unbiased marker sequences for each taxonomic level. Such developments would lead to more accurate abundance of information for genera and species, which underly the establishment of transmission routes (Chapter 4) and causal roles in diseases (Chapter 6). Potential functions, including growth on defined culture media, could be inferred from metagenomic assembled genomes. Variations below the species level, commonly referred to as strains, can also be tracked in the same person overtime, using either metagenomic sequencing or culturomics. With more in vitro characterization and multiomic studies on the commensal microbes, taxonomy according to bioinformatics and metabolic modeling would eventually assimilate traditional taxonomy according to functional measurements.

References

[1] Bowers RM, Kyrpides NC, Stepanauskas R, Harmon-Smith M, Doud D, Reddy TBK, et al. Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. Nat Biotechnol 2017;35:725–31. https://doi.org/10.1038/nbt.3893.

[2] Větrovský T, Baldrian P, Morais D. SEED 2: a user-friendly platform for amplicon high-throughput sequencing data analyses. Bioinformatics 2018;34:2292–4. https://doi.org/10.1093/bioinformatics/bty071.

[3] Prodan A, Tremaroli V, Brolin H, Zwinderman AH, Nieuwdorp M, Levin E. Comparing bioinformatic pipelines for microbial 16S rRNA amplicon sequencing. PLoS One 2020;15. https://doi.org/10.1371/journal.pone.0227434, e0227434.

[4] Sun X, Hu Y-H, Wang J, Fang C, Li J, Han M, et al. Efficient and stable metabarcoding sequencing data using a DNBSEQ-G400 sequencer validated by comprehensive community analyses. Gigabyte 2021;2021:1–15. https://doi.org/10.46471/gigabyte.16.

[5] Qin J, Li Y, Cai Z, Li S, Zhu J, Zhang F, et al. A metagenome-wide association study of gut microbiota in type 2 diabetes. Nature 2012;490:55–60. https://doi.org/10.1038/nature11450.

[6] Sczyrba A, Hofmann P, Belmann P, Koslicki D, Janssen S, Dröge J, et al. Critical assessment of metagenome interpretation—a benchmark of metagenomics software. Nat Methods 2017;14:1063–71. https://doi.org/10.1038/nmeth.4458.

[7] Meyer F, Fritz A, Deng Z-L, Koslicki D, Gurevich A, Robertso G, et al. Critical Assessment of Metagenome Interpretation – the second round of challenges. bioRxiv 2021. https://doi.org/10.1101/2021.07.12.451567.

[8] Almeida A, Nayfach S, Boland M, Strozzi F, Beracochea M, Shi ZJ, et al. A unified catalog of 204,938 reference genomes from the human gut microbiome. Nat Biotechnol 2020. https://doi.org/10.1038/s41587-020-0603-3.

[9] Parks DH, Chuvochina M, Waite DW, Rinke C, Skarshewski A, Chaumeil P-A, et al. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. Nat Biotechnol 2018;36:996–1004. https://doi.org/10.1038/nbt.4229.

[10] Reimer LC, Vetcininova A, Carbasse JS, Söhngen C, Gleim D, Ebeling C, et al. BacDive in 2019: bacterial phenotypic data for high-throughput biodiversity analysis. Nucleic Acids Res 2019;47:D631–6. https://doi.org/10.1093/nar/gky879.
[11] Duran-Pinedo AE, Chen T, Teles R, Starr JR, Wang X, Krishnan K, et al. Community-wide transcriptome of the oral microbiome in subjects with and without periodontitis. ISME J 2014;8:1659–72. https://doi.org/10.1038/ismej.2014.23.
[12] Utter DR, He X, Cavanaugh CM, McLean JS, Bor B. The saccharibacterium TM7x elicits differential responses across its host range. ISME J 2020. https://doi.org/10.1038/s41396-020-00736-6.
[13] Zhu J. Over 50,000 metagenomically assembled draft genomes for the human oral microbiome reveal new taxa and a male-specific bacterium; 2021. p. 2790.
[14] Coleman GA, Davín AA, Mahendrarajah TA, Szánthó LL, Spang A, Hugenholtz P, et al. A rooted phylogeny resolves early bacterial evolution. Science 2021;372. https://doi.org/10.1126/science.abe0511, eabe0511.
[15] Kempes CP, Wang L, Amend JP, Doyle J, Hoehler T. Evolutionary tradeoffs in cellular composition across diverse bacteria. ISME J 2016;10:2145–57. https://doi.org/10.1038/ismej.2016.21.
[16] Mira A, Ochman H, Moran NA. Deletional bias and the evolution of bacterial genomes. Trends Genet 2001;17:589–96. https://doi.org/10.1016/s0168-9525(01)02447-7.
[17] Probst AJ, Auerbach AK, Moissl-Eichinger C. Archaea on human skin. PLoS One 2013;8. https://doi.org/10.1371/journal.pone.0065388, e65388.
[18] Zou Y, Xue W, Luo G, Deng Z, Qin P, Guo R, et al. 1,520 reference genomes from cultivated human gut bacteria enable functional microbiome analyses. Nat Biotechnol 2019;37:179–85. https://doi.org/10.1038/s41587-018-0008-8.
[19] Forster SC, Kumar N, Anonye BO, Almeida A, Viciani E, Stares MD, et al. A human gut bacterial genome and culture collection for precise and efficient metagenomic analysis. Nat Biotechnol 2019;37. https://doi.org/10.1038/s41587-018-0009-7.
[20] Groussin M, Poyet M, Sistiaga A, Kearney SM, Moniz K, Noel M, et al. Elevated rates of horizontal gene transfer in the industrialized human microbiome. Cell 2021;184:2053–2067.e18. https://doi.org/10.1016/j.cell.2021.02.052.
[21] Cross KL, Campbell JH, Balachandran M, Campbell AG, Cooper SJ, Griffen A, et al. Targeted isolation and cultivation of uncultivated bacteria by reverse genomics. Nat Biotechnol 2019. https://doi.org/10.1038/s41587-019-0260-6.
[22] Lagier J-C, Dubourg G, Million M, Cadoret F, Bilen M, Fenollar F, et al. Culturing the human microbiota and culturomics. Nat Rev Microbiol 2018;16:540–50. https://doi.org/10.1038/s41579-018-0041-0.
[23] Truong DT, Franzosa EA, Tickle TL, Scholz M, Weingart G, Pasolli E, et al. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. Nat Methods 2015;12:902–3. https://doi.org/10.1038/nmeth.3589.
[24] Milanese A, Mende DR, Paoli L, Salazar G, Ruscheweyh H-J, Cuenca M, et al. Microbial abundance, activity and population genomic profiling with mOTUs2. Nat Commun 2019;10:1014. https://doi.org/10.1038/s41467-019-08844-4.
[25] Ye SH, Siddle KJ, Park DJ, Sabeti PC. Benchmarking metagenomics tools for taxonomic classification. Cell 2019;178:779–94. https://doi.org/10.1016/j.cell.2019.07.010.
[26] Segata N. MetaPhlAn3; 2021. https://doi.org/10.1101/2020.11.19.388223.
[27] Weyrich LS, Duchene S, Soubrier J, Arriola L, Llamas B, Breen J, et al. Neanderthal behaviour, diet, and disease inferred from ancient DNA in dental calculus. Nature 2017;544:357–61. https://doi.org/10.1038/nature21674.
[28] Wibowo MC, Yang Z, Borry M, Hübner A, Huang KD, Tierney BT, et al. Reconstruction of ancient microbial genomes from the human gut. Nature 2021. https://doi.org/10.1038/s41586-021-03532-0.

[29] Rampelli S, Turroni S, Mallol C, Hernandez C, Galván B, Sistiaga A, et al. Components of a Neanderthal gut microbiome recovered from fecal sediments from El salt. Commun Biol 2021;4:169. https://doi.org/10.1038/s42003-021-01689-y.
[30] Fellows Yates JA, Velsko IM, Aron F, Posth C, Hofman CA, Austin RM, et al. The evolution and changing ecology of the African hominid oral microbiome. Proc Natl Acad Sci 2021;118. https://doi.org/10.1073/pnas.2021655118, e2021655118.
[31] Wood DE, Lu J, Langmead B. Improved metagenomic analysis with kraken 2. Genome Biol 2019;20:257. https://doi.org/10.1186/s13059-019-1891-0.
[32] Cao Y, Lin W, Li H. Large covariance estimation for compositional data via composition-adjusted thresholding. J Am Stat Assoc 2018;1–45. https://doi.org/10.1080/01621459.2018.1442340.
[33] Friedman J, Alm EJ. Inferring correlation networks from genomic survey data. PLoS Comput Biol 2012;8. https://doi.org/10.1371/journal.pcbi.1002687, e1002687.
[34] Scheiman J, Luber JM, Chavkin TA, MacDonald T, Tung A, Pham L-D, et al. Meta-omics analysis of elite athletes identifies a performance-enhancing microbe that functions via lactate metabolism. Nat Med 2019;25:1104–9. https://doi.org/10.1038/s41591-019-0485-4.
[35] Deek RA, Li H. A zero-inflated latent Dirichlet allocation model for microbiome studies. Front Genet 2021;11. https://doi.org/10.3389/fgene.2020.602594.
[36] Liu X, Tong X, Zou Y, Lin X, Zhao H, Tian L, et al. Inter-determination of blood metabolite levels and gut microbiome supported by Mendelian randomization. BioRxiv 2020. https://doi.org/10.1101/2020.06.30.181438. 2020.06.30.
[37] Wang J, Jia H. Metagenome-wide association studies: fine-mining the microbiome. Nat Rev Microbiol 2016;14:508–22. https://doi.org/10.1038/nrmicro.2016.83.
[38] Zhang X, Zhang D, Jia H, Feng Q, Wang D, Di Liang D, et al. The oral and gut microbiomes are perturbed in rheumatoid arthritis and partly normalized after treatment. Nat Med 2015;21:895–905. https://doi.org/10.1038/nm.3914.
[39] Jie Z, Xia H, Zhong S-L, Feng Q, Li S, Liang S, et al. The gut microbiome in atherosclerotic cardiovascular disease. Nat Commun 2017;8:845. https://doi.org/10.1038/s41467-017-00900-1.
[40] Jie Z, Liang S, Ding Q, Li F, Tang S, Wang D, et al. A transomic cohort as a reference point for promoting a healthy gut microbiome. Med Microecol 2021. https://doi.org/10.1016/j.medmic.2021.100039.
[41] Li J, Jia H, Cai X, Zhong H, Feng Q, Sunagawa S, et al. An integrated catalog of reference genes in the human gut microbiome. Nat Biotechnol 2014;32:834–41. https://doi.org/10.1038/nbt.2942.
[42] Browne HP, Forster SC, Anonye BO, Kumar N, Neville BA, Stares MD, et al. Culturing of 'unculturable' human microbiota reveals novel taxa and extensive sporulation. Nature 2016;533:543–6. https://doi.org/10.1038/nature17645.
[43] Darwin C. On the origin of species by means of natural selection, or the preservation of Favoured races in the struggle for life. London: John Murray; 1859.
[44] Mallet J. Darwin and species. In: Ruse M, editor. Cambridge Encycl. Darwin Evol. Thought. Cambridge: Cambridge University Press; 2020. p. 109–15. https://doi.org/10.1017/CBO9781139026895.013.
[45] Wallace AR. The method of organic evolution. Fortn Rev 1895;435–45. NS.57.
[46] Perlmutter JI, Bordenstein SR. Microorganisms in the reproductive tissues of arthropods. Nat Rev Microbiol 2020;18:97–111. https://doi.org/10.1038/s41579-019-0309-z.
[47] Brito IL, Yilmaz S, Huang K, Xu L, Jupiter SD, Jenkins AP, et al. Mobile genes in the human microbiome are structured from global to individual scales. Nature 2016;535:435–9. https://doi.org/10.1038/nature18927.

[48] Fraser C, Alm EJ, Polz MF, Spratt BG, Hanage WP. The bacterial species challenge : ecological diversity. Science 2009;323:741–6.
[49] Sheridan PO, Martin JC, Lawley TD, Browne HP, Harris HMB, Bernalier-Donadille A, et al. Polysaccharide utilization loci and nutritional specialization in a dominant group of butyrate-producing human colonic Firmicutes. Microb Genom 2016;2. https://doi.org/10.1099/mgen.0.000043, e000043.
[50] Rosenthal AZ, Qi Y, Hormoz S, Park J, Li SH-J, Elowitz MB. Metabolic interactions between dynamic bacterial subpopulations. Elife 2018;7. https://doi.org/10.7554/eLife.33099.
[51] Drake JW, Charlesworth B, Charlesworth D, Crow JF. Rates of spontaneous mutation. Genetics 1998;148:1667–86.
[52] Zhao S, Lieberman TD, Poyet M, Kauffman KM, Gibbons SM, Groussin M, et al. Adaptive evolution within gut microbiomes of healthy people. Cell Host Microbe 2019. https://doi.org/10.1016/j.chom.2019.03.007.
[53] Jorth P, Staudinger BJ, Wu X, Hisert KB, Hayden H, Garudathri J, et al. Regional isolation drives bacterial diversification within cystic fibrosis lungs. Cell Host Microbe 2015;18:307–19. https://doi.org/10.1016/j.chom.2015.07.006.
[54] Kent AG, Vill AC, Shi Q, Satlin MJ, Brito IL. Widespread transfer of mobile antibiotic resistance genes within individual gut microbiomes revealed through bacterial hi-C. Nat Commun 2020;11:1–9. https://doi.org/10.1038/s41467-020-18164-7.
[55] Brito IL. Examining horizontal gene transfer in microbial communities. Nat Rev Microbiol 2021;1–12. https://doi.org/10.1038/s41579-021-00534-7.
[56] Baym M, Lieberman TD, Kelsic ED, Chait R, Gross R, Yelin I, et al. Spatiotemporal microbial evolution on antibiotic landscapes. Science 2016;353:1147–51. https://doi.org/10.1126/science.aag0822.
[57] Karr JR, Sanghvi JC, Macklin DN, Gutschow MV, Jacobs JM, Bolival B, et al. A whole-cell computational model predicts phenotype from genotype. Cell 2012;150:389–401. https://doi.org/10.1016/j.cell.2012.05.044.
[58] Covert M. Simultaneous cross-evaluation of heterogeneous e coli datasets via mechanistic simulation. Science 2020. https://doi.org/10.1126/science.eaav3751.
[59] Bordbar A, Monk JM, King ZA, Palsson BO. Constraint-based models predict metabolic and associated cellular functions. Nat Rev Genet 2014;15:107–20. https://doi.org/10.1038/nrg3643.
[60] Mika JT, Poolman B. Macromolecule diffusion and confinement in prokaryotic cells. Curr Opin Biotechnol 2011;22(1):117–26. https://doi.org/10.1016/j.copbio.2010.09.009.
[61] Parry BR, Surovtsev IV, Cabeen MT, O'Hern CS, Dufresne ER, Jacobs-Wagner C. The bacterial cytoplasm has glass-like properties and is fluidized by metabolic activity. Cell 2014;156:183–94. https://doi.org/10.1016/j.cell.2013.11.028.
[62] Shoaie S, Ghaffari P, Kovatcheva-Datchary P, Mardinoglu A, Sen P, Pujos-Guillot E, et al. Quantifying diet-induced metabolic changes of the human gut microbiome. Cell Metab 2015;22:320–31. https://doi.org/10.1016/j.cmet.2015.07.001.
[63] Clark RL, Connors BM, Stevenson DM, Hromada SE, Hamilton JJ, Amador-Noguez D, et al. Design of synthetic human gut microbiome assembly and butyrate production. Nat Commun 2021;12:3254. https://doi.org/10.1038/s41467-021-22938-y.