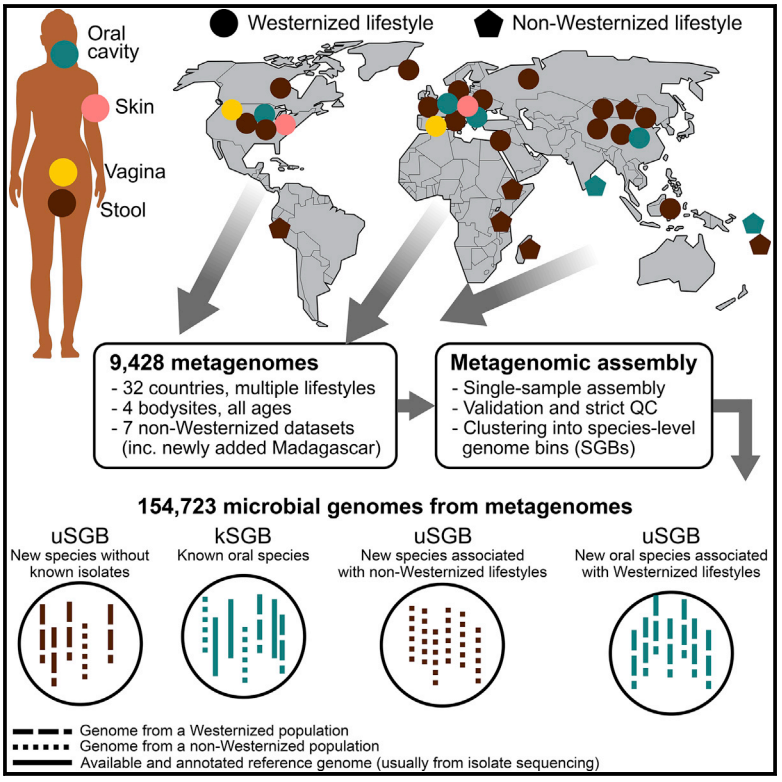
Resource

Cell

Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle

Graphical Abstract



Highlights

- Large-scale metagenomic assembly uncovered thousands of new human microbiome species
- The new genome resource increases the mappability of gut metagenomes over 87%
- Some of the newly discovered species comprise thousands of reconstructed genomes
- Non-Westernized populations harbor a large fraction of the newly discovered species

Authors

Edoardo Pasolli, Francesco Asnicar, Serena Manara, ..., Christopher Quince, Curtis Huttenhower, Nicola Segata

Correspondence

nicola.segata@unitn.it

In Brief

The human microbiome harbors many unidentified species. By large-scale metagenomic assembly of samples from diverse populations, we uncovered >150,000 microbial genomes that are recapitulated in 4,930 species. Many species (77%) were never described before, increase the mappability of metagenomes, and expand our understanding of global body-wide human microbiomes.

Pasolli et al., 2019, Cell *176*, 649–662
January 24, 2019
https://doi.org/10.1016/j.cell.2019.01.001

CellPress

Resource

Cell

Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle

Edoardo Pasolli,[1] Francesco Asnicar,[1,8] Serena Manara,[1,8] Moreno Zolfo,[1,8] Nicolai Karcher,[1] Federica Armanini,[1] Francesco Beghini,[1] Paolo Manghi,[1] Adrian Tett,[1] Paolo Ghensi,[1] Maria Carmen Collado,[2] Benjamin L. Rice,[3] Casey DuLong,[4] Xochitl C. Morgan,[5] Christopher D. Golden,[4] Christopher Quince,[6] Curtis Huttenhower,[4,7] and Nicola Segata[1,9,*]

[1]CIBIO Department, University of Trento, Trento, Italy
[2]Institute of Agrochemistry and Food Technology-National Research Council, Valencia, Spain
[3]Harvard University, Cambridge, MA, USA
[4]Harvard T.H. Chan School of Public Health, Boston, MA, USA
[5]University of Otago, Otago, New Zealand
[6]Warwick Medical School, University of Warwick, Warwick, UK
[7]The Broad Institute, Cambridge, MA, USA
[8]These authors contributed equally
[9]Lead Contact
*Correspondence: nicola.segata@unitn.it
https://doi.org/10.1016/j.cell.2019.01.001

SUMMARY

The body-wide human microbiome plays a role in health, but its full diversity remains uncharacterized, particularly outside of the gut and in international populations. We leveraged 9,428 metagenomes to reconstruct 154,723 microbial genomes (45% of high quality) spanning body sites, ages, countries, and lifestyles. We recapitulated 4,930 species-level genome bins (SGBs), 77% without genomes in public repositories (unknown SGBs [uSGBs]). uSGBs are prevalent (in 93% of well-assembled samples), expand underrepresented phyla, and are enriched in non-Westernized populations (40% of the total SGBs). We annotated 2.85 M genes in SGBs, many associated with conditions including infant development (94,000) or Westernization (106,000). SGBs and uSGBs permit deeper microbiome analyses and increase the average mappability of metagenomic reads from 67.76% to 87.51% in the gut (median 94.26%) and 65.14% to 82.34% in the mouth. We thus identify thousands of microbial genomes from yet-to-be-named species, expand the pangenomes of human-associated microbes, and allow better exploitation of metagenomic technologies.

INTRODUCTION

Despite extensive recent studies of the human microbiome using a variety of culture-independent molecular technologies (Human Microbiome Project Consortium, 2012; Qin et al., 2010; Quince et al., 2017a; Rinke et al., 2013), most characterization of these ecosystems is still focused on microbes that are easily cultivable, particularly when those with sequenced isolate genomes are considered. Since physiological characterization of diverse, uncharacterized human-associated microbes by cultivation can be difficult in high throughput (Browne et al., 2016), additional approaches are needed that scale with the extent of populations that can now be surveyed using metagenomic sequencing. Culture-independent genomic approaches that are scalable to large cohorts (Human Microbiome Project Consortium, 2012; Qin et al., 2010; Quince et al., 2017a) have facilitated access to an expanded set of isolation-recalcitrant members of the microbiome, but they also suggested the presence of a large fraction of still unexplored diversity (Nielsen et al., 2014; Rinke et al., 2013).

Here, we present a set of 154,723 microbial genomes that are often prevalent, population specific, and/or geographically specific that we reconstructed via single-sample assembly from a total of 9,428 global, body-wide metagenomes. Other studies have also succeeded in reconstructing microbial genomes by metagenomic assembly on single human cohorts (Bäckhed et al., 2015; Brooks et al., 2017; Ferretti et al., 2018; Human Microbiome Project Consortium, 2012; Raveh-Sadka et al., 2015; Sharon et al., 2013), but systematic cross-study cataloging of metagenomically assembled genomes focused so far on non-human environments (Oyama et al., 2017; Parks et al., 2017). Complementary techniques, such as co-abundance of gene groups (Nielsen et al., 2014), can identify genomic bins without reference, but these techniques do not account for sample-specific strains and strain-level differences in the sequence reconstruction and thus require downstream single-nucleotide variation analysis on specific genomic regions to uncover strain variability (Quince et al., 2017b; Truong et al., 2017).

Using large-scale single-sample metagenomic assembly supported by strict quality control (including filtering based on nucleotide polymorphisms), we identified 3,796 species-level clades (comprising 34,205 genomes) without previous whole-genome

information. This identified several taxa prevalent but previously unobserved even in well-profiled populations (e.g., a genus-level Ruminococcaceae clade phylogenetically close to *Faecalibacterium*), extensive taxonomically uncharacterized species associated with non-Western populations, and the presence of several taxa from undersampled phyla (e.g., Saccharibacteria and Elusimicrobia) in oral and gut microbiomes. The resulting genome set can thus serve as the basis for future strain-specific comparative genomics to associate variants in the human microbiome with environmental exposures and health outcomes across the globe.

RESULTS

Recovering Over 150,000 Microbial Genomes from ~10,000 Human Metagenomes

We employed a very large-scale metagenomic assembly approach to reconstruct bacterial and archaeal genomes populating the human microbiome (see STAR Methods). From a total of 9,316 metagenomes spanning 46 datasets from multiple populations, body sites, and host ages (Table S1), and an additional cohort from Madagascar (Golden et al., 2017) (STAR Methods; Table S1), we reconstructed a total of 154,723 genomes (each made up of a group of clustered contigs; see STAR Methods) using a single-sample assembly strategy tailored at maximizing the quality rather than the quantity of genomes reconstructed from each sample. The resulting catalog greatly expands the set of ~150,000 microbial genomes publicly available (see STAR Methods). All assembled genomes passed strict quality control including estimation of completeness, contamination, and a measure of strain heterogeneity (see STAR Methods), and they exceed the thresholds to be defined medium quality (MQ) according to recent guidelines (Bowers et al., 2017) (completeness >50%, contamination <5%). The quality of these genomes was comparable with that of isolate sequencing (STAR Methods; Table S2) and in line also with the quality achievable by manually curated metagenomic approaches (Table S2) and time-series or cross-sectional metagenomic co-binning (see STAR Methods; Table S2). Genomes may include contigs from plasmids (see STAR Methods), and stricter quality control reduced the set of near-complete, high-quality (HQ) genomes to 70,178 with completeness higher than 90% and reduced probability of intra-sample strain heterogeneity (<0.5% polymorphic positions, see STAR Methods). The main characteristics of HQ genomes are in line and in some cases better than those from the compendium of reference genomes available in public repositories, although MQ genomes also had similar quality scores compared to HQ genomes (modulo completeness; STAR Methods). The set of genomes we reconstructed (Table S3; Data and Software Availability) and the associated 2.85 million (M) total functional annotations (STAR Methods; Figure S1) are thus appropriate as a basis for more in-depth microbial community analyses.

Human Microbiome Genomes Belong to ~5,000 Functionally Annotated SGBs

To organize the 154,723 genomes into species-level genome bins (SGBs), we employed an all-versus-all genetic distance quantification followed by clustering and identification of genome bins spanning a 5% genetic diversity, which is consistent with the definition of known species (see STAR Methods) and with other reports (Jain et al., 2018). We obtained 4,930 SGBs from 22 known phyla (Figure 1A; Table S4). This is likely an underestimate of the total phylum-level diversity, because some SGBs are very divergent from all previously available reference genomes and cannot be confidently assigned to a taxonomic family (Table S4): 345 SGBs (58% of which with HQ or multiple reconstructed genomes) display more than 30% Mash-estimated genetic distance (Ondov et al., 2016) from the closest isolate with a phylum assignment (Figure S2A). The SGB genomic catalog spans on average 3.0%, SD 1.8% intra-SGB nucleotide genetic variability, and each SGB contains up to 3,457 genomes from different individuals (average 31.4, SD 147.6; Figures 1C and S2B).

Functional annotation of all the reconstructed genomes assigned a UniRef90 (The UniProt Consortium, 2017) label to 230 M genes and a UniRef50 to 268 M genes (72.7% and 84.8% of the total of 316 M genes, respectively). Additional EggNOG (Huerta-Cepas et al., 2017) labels were assigned to 80.8% of the 4,930 SGBs' genome representatives. The functional potential profiles of the genomes had, as expected, clear phylogenetic differentiation (Figure S1), and the rate of annotation varied greatly in SGBs (e.g., >90% genes annotated for well-studies species such as *Escherichia coli* or *Bacteroides fragilis* versus 22% for ID 15286, which is the largest SGB without reference genomes). Each of the body sites considered had a clear distinctive set of annotations with the adult fecal microbiome enriched for 101,056 gene families (Table S5, Bonferroni-corrected Fisher's test $p < 0.01$), representative of anaerobe-specific functions such as formate oxidation and methanogenesis and a strong representation of biofilm formation functions in the oral cavity and on the skin. Genomes from the stool microbiome of newborns had 94,562 enriched gene families (Table S5, Bonferroni-corrected Fisher's test $p < 0.01$) comprising a variety of functions such as folate biosynthesis and lactose, oligosaccharides, and mucin degradation that are typical of the niche and nutritional regime of unweaned infants (Asnicar et al., 2017; Marcobal et al., 2011; Yatsunenko et al., 2012). Age-specific functions (Table S5) are characterized by the later host developmental stages of children (17,121 specific functions) and school-age individuals (349 specific functions). The Westernization process has also a strong influence on the functions encoded in the stool microbiome, with a total of 106,872 differential families (Table S5, Bonferroni-corrected Fisher's test $p < 0.01$) spanning enzymes involved in the metabolism of complex carbohydrates, such as xylose and cellulose, and in specific cobalamin biosynthesis pathways; these are likely reflecting dietary habits, among other environmental differences. The organization of the reconstructed genomes in SGBs and their functional profiling will be the basis for comprehensive future metagenomic characterizations.

The Reconstructed Genomes and SGBs Increase the Diversity and Mappability of the Human Microbiome

We identified 3,796 SGBs (i.e., 77.0% of the total) covering unexplored microbial diversity as they represent species without any publicly available genomes from isolate sequencing or previous metagenomic assemblies (Figures 1B and S3A). These SGBs, that we named unknown SGBs (uSGBs), include on average

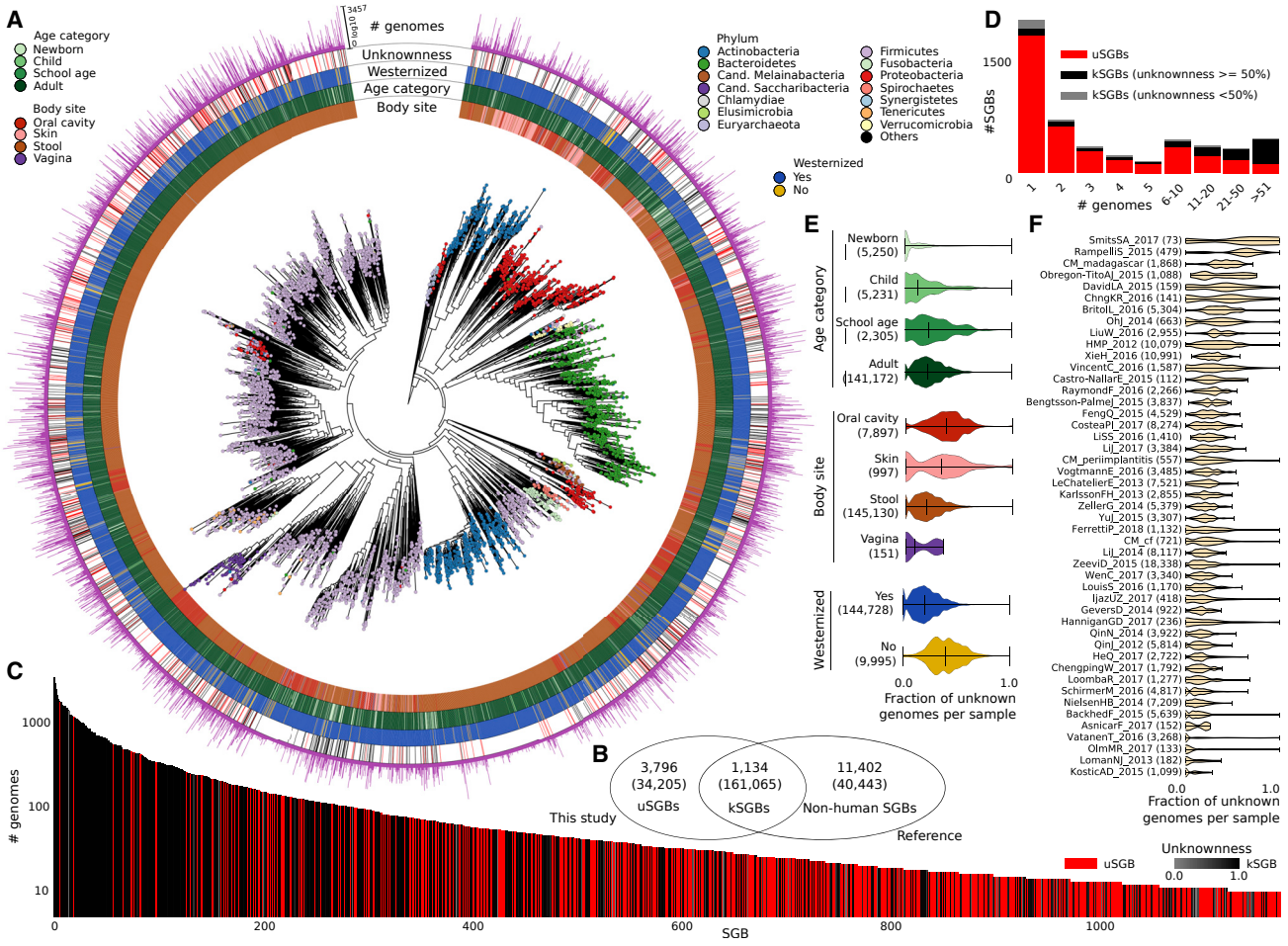


Figure 1. 4,930 SGBs Assembled from 9,428 Meta-analyzed Body-wide Metagenomes

(A) A human-associated microbial phylogeny of representative genomes from each species-level genome bin (SGB). Figure S3A reports the same phylogeny but including isolate genomes not found in the human-associated metagenomes.

(B) Overlap of SGBs containing both existing microbial genomes (including other metagenomic assemblies) and genomes reconstructed here (kSGBs), SGBs with only genomes reconstructed here and without existing isolate or metagenomically assembled genomes (uSGBs), and SGBs with only existing genomes and no genomes from our metagenomic assembly of human microbiomes (non-human SGBs).

(C) Many SGBs contain no genomes from sequenced isolates or publicly available metagenomic assemblies (uSGBs). Only SGBs containing >10 genomes are shown.

(D) Fraction of uSGBs and kSGBs as a function of the size of the SGBs (i.e., number of genomes in the SGB).

(E) Distribution of the fraction of uSGBs in each sample by age category, body site, and lifestyle.

(F) Distribution of the fraction of uSGBs in each study.

9.0, SD 45.4 reconstructed genomes, and 1,693 of them (45%) had at least one HQ genome. Recursive clustering of SGBs' representatives at genus- and family-level genetic divergence (see STAR Methods) provided taxonomic context for 75.2% of the uSGBs with 1,472 assignments to genera and 1,383 more to families (Table S4). The 941 uSGBs that were left unplaced at family level remained unassigned for limitations of whole-genome similarity estimates, but we report the similarity and taxonomy of the closest matching strain (Table S4).

Only 1,134 of the 4,930 SGBs represent at least partially known SGBs (kSGBs) that include one or more genomes in public databases. This number of kSGBs is consistent with the 1,266 species we found at least once in the same set of metagenomes (Pasolli et al., 2017) at >0.01% abundance using reference-based taxonomic profiling (Truong et al., 2015). Most uSGBs represent instead relatively rare human-associated microbes (46.7% of uSGBs comprise one reconstructed genome only, Table S4, and 46.1% genomes in uSGBs are at <0.5% relative abundance, STAR Methods and Table S4), but some uSGBs are highly prevalent, with 10 uSGBs in the set of the 100 SGBs with the largest number of reconstructed genomes (Figures 1C, 1D, and S2B) and 368 genomes in uSGBs accounting for >10% of reads. Because many uSGBs are associated with specific sample types (e.g., oral cavity or non-Westernized samples, Figure 1E), the actual number of possibly redundant genomes they contain is likely underestimated for those sample

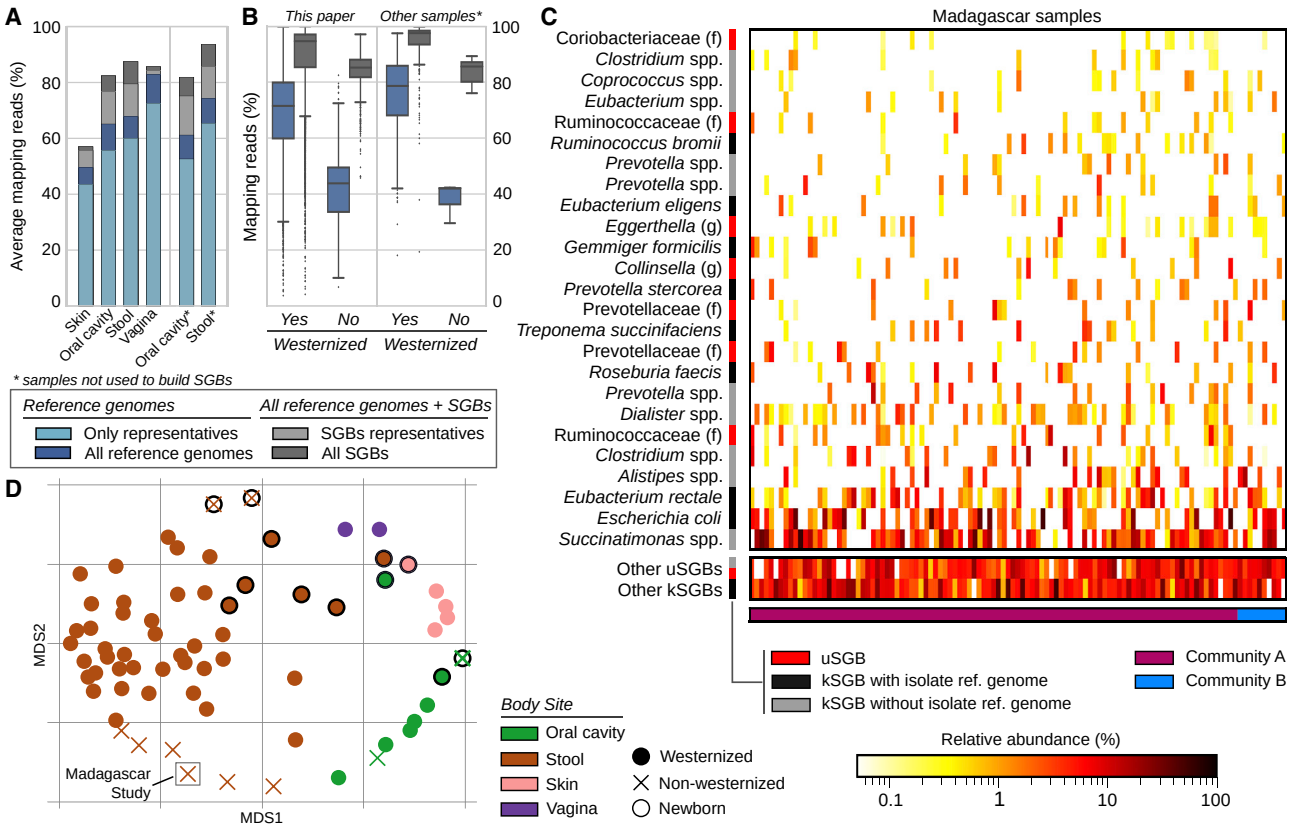


Figure 2. The Expanded Genome Set Substantially Increases the Mappability of Human Metagenomes

(A) We mapped the subsampled original 9,428 metagenomes and 389 additional samples not considered for building the SGBs against the 154,723 reconstructed genomes and 80,990 previously available genomes. Raw-read mappability increased significantly (Mann-Whitney U test, $p < 1e{-}50$), e.g., from an average of 67.76% to 87.51% in the gut. Representative genomes refer to the highest-quality genomes selected from the 4,930 human SGBs and the 11,402 non-human SGBs. Extended statistics are in Figure S4.

(B) Metagenomic read mappability increases more in non-Westernized than Westernized gut microbiomes (Welch's t test, $p < 1e{-}50$), both when considering samples used for SGBs' reconstruction (26.50% average increase in 7,059 Westernized samples versus 96.56% in 454 non-Westernized samples) and when considering 264 additional samples not used for SGBs' reconstruction (25.16% versus 117.40% average increase, respectively).

(C) The gut microbiomes from Madagascar we sequenced here showed several highly abundant uSGBs and a large set of SGBs reconstructed in only subsets of the samples. Many kSGBs in this dataset do not contain isolate genomes but only previous metagenomic assemblies. The 25 most abundant SGBs are reported and ordered according to their average relative abundance.

(D) Multidimensional scaling on datasets using the Bray-Curtis distance on per-dataset SGB prevalences highlights distinct microbial communities between Westernized and non-Westernized populations within and between body sites and age categories.

types with comparably fewer metagenomes available. Functional annotation of uSGB genomes assigned a UniRef90 cluster to only 31.9% of the genes, while the annotation rate increased to 81.0% for kSGB genomes.

The expanded human microbiome diversity induced by the uSGBs (200% increase in the reconstructed phylogenetic branch length, 50% considering only uSGBs with >10 genomes, Figure 1A) can be crucial as a genomic reference in the characterization ("mappability") of the sequence information in a metagenome. Genomes in uSGBs are indeed responsible for a substantial decrease of the metagenomic reads that do not match any microbial reference (Figures 2A and S4). This is due both to uSGBs representing target microbes without assigned species (16.76% average increase using only representative genomes of uSGBs, Figure 2A) and to the expansion of pangenomes of kSGBs and uSGBs (27.84% increase when considering all genomes instead of only SGB representatives). On average, the read mappability for stool samples reached 87.51% (29.14% increase, Figure 2A) and 82.34% in the oral cavity (26.40% increase, Figure 2A). Some outlier samples decreased the averages as the median final mappabilities were higher, reaching 94.26% for the stool microbiome and 90.13% for the oral microbiome in Westernized populations. The mappability of the skin microbiome was also increased (15.17% increase) but reached a lower overall value (57.07%) because fewer skin samples were available and non-bacterial organisms such as the molluscum contagiosum virus (Oh et al., 2014) and fungi from the *Malassezia* genus (Tett et al., 2017) also populate the skin. Mappability in the vaginal microbiomes was instead already high (82.77%) due to a reduced panel of known species dominating the large majority of these communities, but the set of 4,930 reconstructed SGBs still increased the mappability by

3.42%. The mappability increase is dramatic for the gut microbiomes of non-Westernized populations that are very poorly represented by available reference genomes (42.33% mappability) and can now reach a mappability of 83.20%, which is comparable with that of Westernized populations (Figure 2B). These substantial gains in read mappability when using our genome catalog are achieved also for stool and oral samples not used to construct the resource (STAR Methods; Figures 2A and 2B), confirming its relevance as reference for future studies.

SGBs without publicly available genomes (uSGBs) represent 34,205 reconstructed genomes (Figure 1B), belonging to metagenomes in different body sites, ages, and general lifestyles (Figures 1E and 1F). Microbiomes with lower diversity, such as those from infants or the female urogenital tract, carried a generally lower fraction of uSGBs. Populations with non-Westernized lifestyles—including the Madagascar cohort we sequenced (Figures 2C and 2D)—conversely yielded a fraction of genomes in uSGBs nearly double that of Western-style populations (average 40% and 21%, respectively, $p < 1e{-}50$, Figure 1E). Most of the abundant kSGBs in the Madagascar cohort do not include isolate genomes but only sequences from previous metagenomic assemblies (Figure 2C), and these uSGBs and poorly characterized kSGBs are contributing to the clear distinction of the gut microbiome with respect to general lifestyles (Figure 2D). The higher rate of uSGB recovery in non-Westernized populations is likely the consequence of comparatively fewer studies profiling these populations and their more diverse gut microbiomes.

The Diversity of Human-Associated Archaea and Bacterial Phyla Is Expanded by uSGBs

Many clades, including some phyla, were greatly expanded by reconstructed genomes belonging to species that do not have deposited genome sequences or taxonomic labels (uSGBs). For example, the candidate phylum Saccharibacteria (previously named TM7) contains members of the oral microbiome that are particularly difficult to cultivate (He et al., 2015; Solden et al., 2016). For this clade, we reconstructed 387 genomes from 108 SGBs (Figure 1A), some representing members observed only using 16S rRNA gene sequencing (Brinig et al., 2003; Segata et al., 2012a). An isolate reference genome was only available for a single SGB within this clade (ID 19849); the other 16 reference genomes for this phylum were undetected in oral cavity metagenomes (Figure S3B). The 107 Saccharibacteria uSGBs thus suggest a substantially undersampled diversity of human-associated members of this phylum. Its importance is also confirmed by the occurrence of at least one genome from these 108 SGBs in 33% of oral cavity samples, where they can reach average abundances above 3% (Table S4) and maximum abundances exceeding 10%.

We further recovered 675 genomes of Archaea (526 from 6 kSGBs and 149 from 13 uSGBs, Figure 1A) and reconstructed its phylogeny (Figure S3C). More than half of these genomes (n = 487) belonged to the *Methanobrevibacter smithii* kSGB (ID 714), which was present at relatively low abundance (average 1.06%, SD 1.26%). A related but diverged SGB including 94 genomes was identified (ID 713, 5.6% nucleotide divergence from the *M. smithii* isolate genome) at comparable abundance (average 0.92%, SD 2.02%), but it notably accounted for up to 20% of all reads in some gut samples. Among uSGBs, we also reconstructed genomes assigned to *Thermoplasmatales* (ID 376, 378, 380, 381), Candidatus *Methanomethylophilus* (ID 372, 382, 384), *Methanomassiliicoccus* (ID 362, 364), and *Methanosphaera* (ID 697), all very distant from their nearest reference genomes (average 22.4%, SD 4.0% nucleotide distance). This expanded human-associated archaeal diversity suggests the presence of several as-yet-uncharacterized archaea of potentially unique functional relevance in this ecosystem.

Several Prevalent Uncharacterized Intestinal Clostridiales Clades Occur Phylogenetically between Ruminococcus and Faecalibacterium

Some of the uSGBs with the largest number of reconstructed genomes are also highly abundant in the gut microbiome, with 1,153 uSGBs totaling >13,000 genomes each present in the sample where it has been reconstructed at an average abundance >1% (and 172 uSGBs at >5% average abundance). Among them, uSGB ID 15286, that we named "*Candidatus* Cibiobacter qucibialis", is the most prevalent uSGB, comprising 1,813 reconstructed genomes. This species is phylogenetically placed between *Faecalibacterium* and *Ruminococcus* (Figures 3A and S5A), key members of the gut microbiome that are typically present at comparably lower abundances (1.84% *Faecalibacterium* kSGB and 1.29% *Ruminococcus* kSBG in contrast to 2.47% *Ca.* Cibiobacter qucibialis). Six other prevalent (1,563 total genomes) and abundant (1.14% average abundance) SGBs occurred monophyletically in the same subtree between faecalibacteria and ruminococci (Figure 3A). Only one of these seven total SGBs contains an isolate genome, which is the recently sequenced *Gemmiger formicilis* genome (Gossling and Moore, 1975) included in kSGB ID 15300 (1,212 genomes, Figures 3A and 3B). A genome from the *Subdoligranulum variabile* species, itself not found in any of the study's assemblies, was the only other reference phylogenetically close to this clade, explaining the previous identification of an unknown *Subdoligranulum* ("*Subdoligranulum* unclassified") as the most prevalent single taxon in reference-based profiles of the gut microbiome (Pasolli et al., 2017). This prevalent 7-SGBs clade comprising 3,370 reconstructed genomes that can be very abundant (>5% relative abundance in >200 samples) is thus an important but so far neglected genus-level lineage in the human microbiome.

In an estimated maximum-likelihood whole-genome phylogeny of the 1,813 genomes belonging to *Ca.* Cibiobacter qucibialis (Figure 3C), genomes of non-Westernized populations were placed together in a monophyletic subtree (Figure 3C). This subtree included 26 strains from the Madagascar microbiomes we sequenced in this work, in addition to strains from three other populations with traditional lifestyles but differing geographic locations (Figure 3D). Although the non-Westernized subtree includes few genomes (2% of the total), this is a consequence of limited sampling from these population types because the prevalence of this SGB in Westernized populations is comparable (23% against 15% in non-Westernized populations). No clear internal clustering was evident for Westernized samples (Figure 3C), except for a large set of 222 samples retrieved from

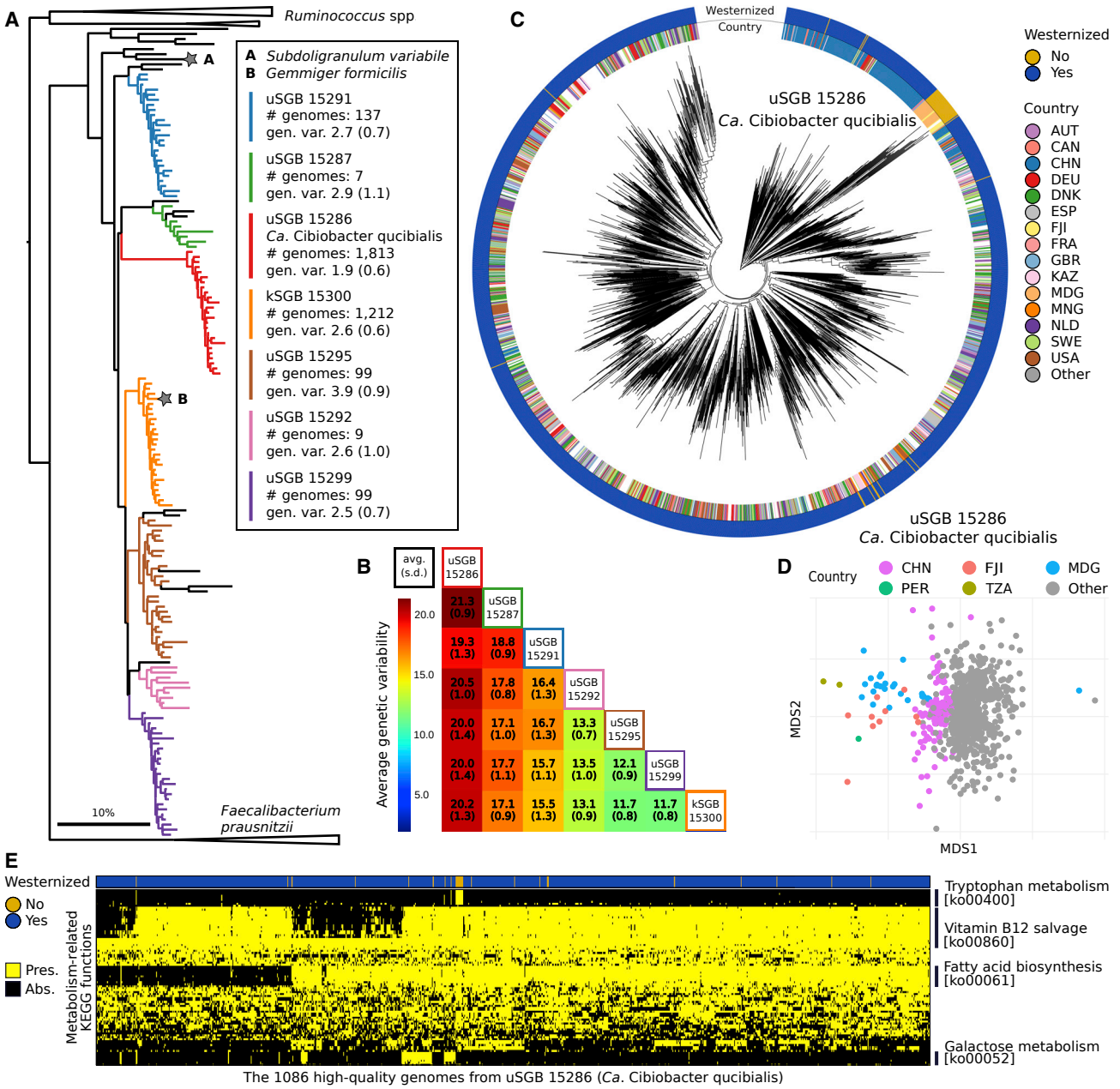


Figure 3. Several Prevalent Intestinal uSGBs Are Found within the Clostridiales Order Related to *Ruminococcus* and *Faecalibacterium*

(A) All SGBs in the assembled phylogeny (Figure 1A) placed between reference genomes for *Ruminococcus* and *Faecalibacterium* species that are reported as collapsed trees. A maximum of 25 HQ genomes from each SGB are displayed, and SGBs with <3 genomes are left black.

(B) The monophyletic clade with the six uSGBs and the kSGB containing *Gemmiger formicilis* represent clearly divergent species with inter-species genetic distance typical of genus-level divergence (average 16.6%, SD 3.1% nucleotide distance).

(C) A whole-genome phylogeny for the 1,806 genomes in *Ca.* Cibiobacter qucibialis (STAR Methods). Some subtrees associate with geography and non-Westernized populations, while others seems to be geography- and lifestyle-independent (see text).

(D) Multidimensional scaling of genetic distances among genomes of *Ca.* Cibiobacter qucibialis highlights the divergence of strains carried by non-Westernized populations, with Chinese populations subclustering within the large cluster of Westernized populations.

(E) Madagascar-associated strains of *Ca.* Cibiobacter qucibialis (uSGB 15286) uniquely possess the *trp* operon for tryptophan metabolism (Table S7). Other functional clusters in Westernized strains from geographically heterogeneous populations include vitamin B12 and fatty acid biosynthesis and galactose metabolism. The KEGG functions present in >80% or in <20% of the samples were discarded except for significant associations with lifestyle.

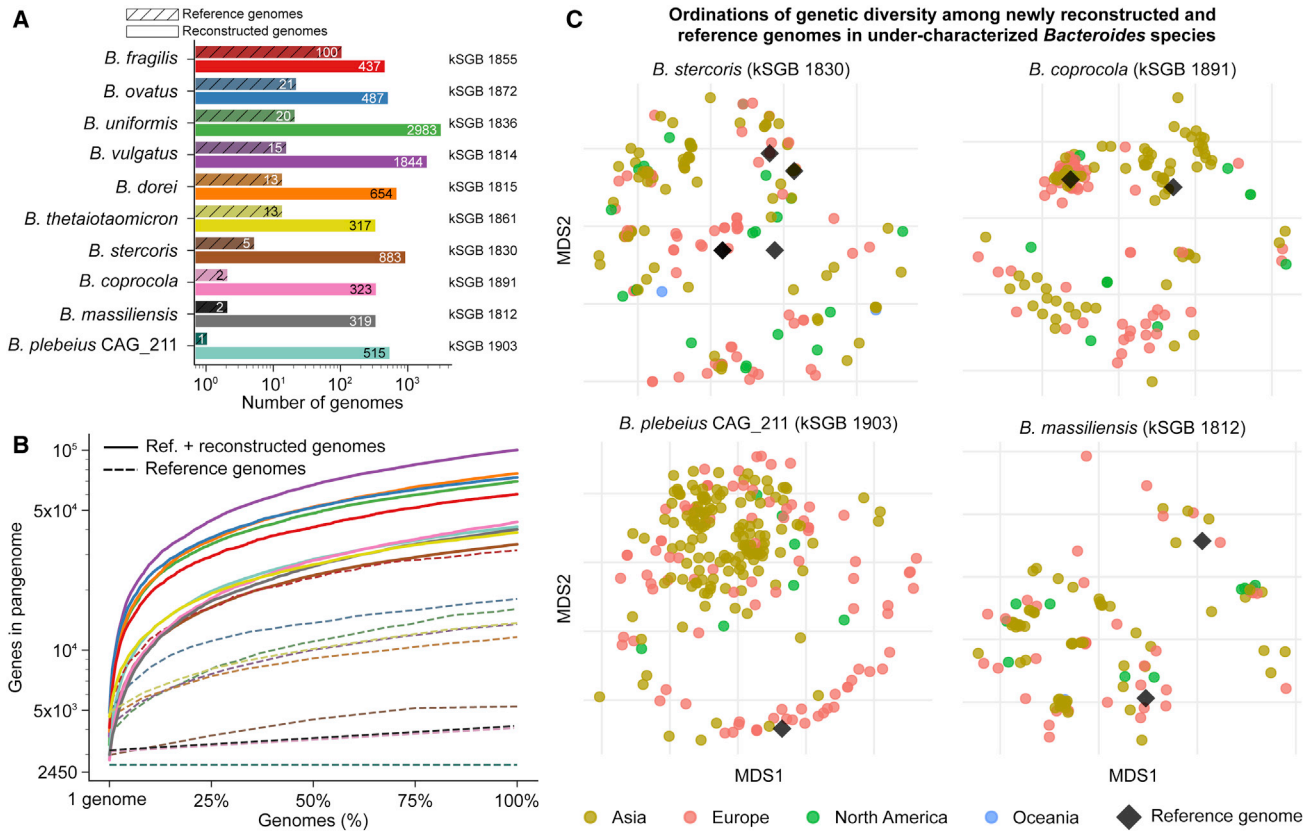


Figure 4. The Metagenomically Reconstructed Genomes Greatly Expand the Genetic and Functional Diversity of the Ten *Bacteroides* Species Most Prevalent in the Human Gut

(A) Additional *Bacteroides* genomes we assembled from metagenomes increase the size of the ten most prevalent *Bacteroides* kSGBs from 4 to >500 times.

(B) The expanded *Bacteroides* kSGBs account for much larger pangenomes that capture a greater functional potential.

(C) Ordinations on intra-SGB genetic distances (fractions of nucleotide mutations in the core genome) highlight the genetic structure of *Bacteroides* species and that reference genomes were available only for a reduced subset of subspecies structures (additional ordinations are in Figure S6A).

the seven Chinese cohorts that are monophyletically placed in the same subtree despite widely different pre-sequencing protocols (Table S6) and resemble non-Westernized genomes (Figures 3C and 3D). This suggests a complex process of gut microbial ecological establishment in which both host lifestyle and biogeography play roles with comparable effect sizes.

Functional potential profiling of SGBs can suggest metabolic features that distinguish each clade, and for *Ca.* Cibiobacter qucibialis, we found functional modules specific to only some of the constituent strains (Figure 3E; Table S7). These include the pathway for the biosynthesis of vitamin B12 from precorrin-2, lacking in some Westernized strains that instead use other pathways for vitamin B12 production, as well as gene clusters devoted to fatty acid biosynthesis and galactose metabolism (Figure 3E). A strong lifestyle-associated difference characterized the non-Westernized strains in *Ca.* Cibiobacter qucibialis (uSGB 15286), as they were the only strains in this SGB with the whole set of genes in the *trp* operon for tryptophan metabolism. The Trp biosynthetic pathway can be organized as whole-pathway operon or as dispersed genes in different bacterial species (Merino et al., 2008), as a result of organismal divergence, adjustment to environmental availability of key molecules, and lateral gene transfer events (Xie et al., 2003). We speculate that the presence of the whole operon in the non-Westernized strains may be indicative of divergent evolution in the Westernized strains of *Ca.* Cibiobacter qucibialis, potentially as a consequence of a loss-of-operon event.

Sample-Specific Strain Recovery Greatly Enlarges the Pangenomes of Key Intestinal Microbes

Bacteroides are among the most studied intestinal species (Marcobal et al., 2011) and are core in European and American populations (Human Microbiome Project Consortium, 2012; Nielsen et al., 2014), but our analysis still recovered unsampled intra-species diversity. Among the ten largest SGBs, the number of available isolate genomes ranges from 1 (*Bacteroides plebeius*) to 100 (*Bacteroides fragilis*), whereas we added from 317 to 2,983 individual representatives (Figure 4A). These expanded genome sets provide much larger collections of distinct genes that can be present in strains of each species, i.e., pangenomes, which spanned ~30,000 to >70,000 genes per *Bacteroides* species, capturing a substantially wider functional potential

compared to isolate genomes (Figure 4B). The number of genomes in a species bin did not correlate well with the size of the associated pangenome (Pearson correlation 0.48, p = 0.16), indicating that pangenome recovery is not simply a function of the amount of associated sequence. No *Bacteroides* pangenomes approached saturation even given the amount of sequence included in this study (average of 276, SD 93 additional pan-genes when moving from the 99th percentile to the whole set of reconstructed genomes), suggesting that even for common, well-studied organisms, a surprising amount of intra-species genomic diversity (and associated biochemical function) remains to be captured.

Most of the *Bacteroides* SGBs contained distinct subspecies clusters, and many of these subspecies include only genomes we reconstructed in this work (Figures 4C and S6A). Some of the most abundant *Bacteroides* species (including *B. stercoris* and *B. plebeius*) were only partially captured by isolate genomes, and the additional reconstructed genomes accounted for an average of 95.8%, SD 5.0% total branch length in the ten core-genome phylogenies. Considering that genetic sub-speciation is highly correlated with functional diversification (correlations > 0.8, p < 1e−50, Figure S6B), the reconstructed genomes thus uncover not only genetic diversity but also relevant functional diversity included in otherwise inaccessible *Bacteroides* subspecies.

Some uSGBs and Subspecies Are Strongly Associated with Non-Westernized Populations

To further assess the specificity of the unexplored uSGBs among global populations, we profiled the gut microbiomes of two rural communities with non-Western lifestyles from northeastern Madagascar (STAR Methods). The SGB profiles of the Madagascar population were profoundly different from that of Western-style populations (Figures 2C and 2D), with 49 of the 941 large (>10 genomes) SGBs highly enriched in this east-African population and 8 SGBs uniformly absent (20 total depleted SGBs, Fisher's test Bonferroni-corrected p < 0.05, Figure 5A, Table S6). An SGB that contains a previously co-assembled *Succinatimonas* sp. but no isolate genomes was the strongest association with the Madagascar population (Fisher's Bonferroni-corrected p = 8.2e−99), as well as with non-Westernized populations generally (p = 4.3e−244), across which it was successfully assembled in 55.9% of the samples (4.55% average and 56% maximum relative abundance) compared to only 1.6% in Westernized samples (3.34% average and 20.13% maximum relative abundance). The type strain of this genus (*Succinatimonas hippei*) was isolated from the gut of a healthy Japanese individual in 2010 (Morotomi et al., 2010) and is phylogenetically similar to isolates from poultry. The ability to degrade D-xylose is characteristic of the clade, a plant-sugar whose metabolism was previously reported as enriched in rural microbiomes (De Filippo et al., 2010). The phylogenetic structure of *Succinatimonas* SGB 3677 also suggests further specialization to specific host lifestyles at the subspecies level, with 99 of the 117 genomes from Westernized populations tightly clustering together and well separated from all 246 genomes from the five non-Westernized populations (Figures 5C and S5B). This SGB in the *Succinatimonas* genus shows a geographically consistent pattern of lifestyle association, resulting in dramatically different prevalences across the globe (p = 4.3e−244) as well as intra-species geographically specific genetic diversification.

The non-Westernized gut microbiome is overall enriched for uSGBs rather than kSGBs (Figures 5A and 5B), which was consistent despite the different protocols used in the considered studies (Table S6). These include several uSGBs in the Firmicutes and Actinobacteria phyla but also in less typically human-associated phyla such as the Elusimicrobia phylum. Two Elusimicrobia uSGBs were associated with the Madagascar (ID 19692 and ID 19694, Fisher's test p = 4.64e−11 and 9.76e−05, respectively) and non-Westernized gut microbiome (ID 19694, p = 1.52e−53) but showed 22% nucleotide divergence from the closest isolate genome (Figures 5D and S5A). 22 isolate genomes are available for this phylum, but they were typically recovered from termites and other insects (Herlemann et al., 2007) and were even more genetically distant from those we identified in humans (>30% nucleotide distance). While these divergent Elusimicrobia uSGBs populate the non-Westernized gut microbiome with some frequency (15.4% prevalence, 0.73% average relative abundance, Figure 5D), they are rarely found in Westernized individuals (0.31% prevalence).

Bacteroides uniformis was the strongest Westernized-lifestyle-associated bacterium (Figure 5B; Table S6), and 13 other *Bacteroides* species with a combined total of 10,992 genomes also showed the same trend (2.66% versus 0.86% prevalence and 5.77% Westernized versus 1.69% non-Westernized average abundance). With the exception of four unnamed low-prevalence *Bacteroides* SGBs (434 genomes in total), no species of this clade was significantly enriched in non-Westernized populations; instead, these were highly enriched in *Prevotella* species (12 kSGBs against no significant *Prevotella* kSGB in Westernized populations), as expected (De Filippo et al., 2010; Obregon-Tito et al., 2015). Several other known and relatively well-characterized species (including *Alistipes putredinis*, *Parabacteroides distasonis*, and *Akkermansia muciniphila*) were significantly associated with Westernized populations, in total accounting for >23 times more kSGBs than uSGBs. Conversely, among SGBs enriched in non-Westernized populations, uSGBs greatly outnumbered kSGBs (144 versus 63, Fisher's test p = 1.0e−23). This further confirms that populations with non-urbanized and traditional lifestyles have a more uncharacterized gut microbiome that is made more accessible to future characterization by these results.

Microbiome differentiation between lifestyles was also reflected at the functional level (Figure 5E; Tables S5 and S6). Sulfur energy metabolism (ko00920), vitamin B12 salvage (ko00860), and the sodium-ion-specific ATP synthase operon *ntp* (ko00190) were among the KEGG functional modules significantly enriched in Westernized microbiomes (Figure 5E). Other functions were present in both lifestyles but encoded by different enzymes and pathways. For example, both groups' microbiomes encoded extensive antibiotic biosynthesis genes (Figure 5E), but while Westernized-enriched SGBs encoded the pathway for penicillin and cephalosporin biosynthesis (ko00311), non-Westernized-enriched SGBs more often carried genes for macrolide biosynthesis (ko00523). Similarly, genes for tryptophan metabolism were differently present in the two groups, with parts of

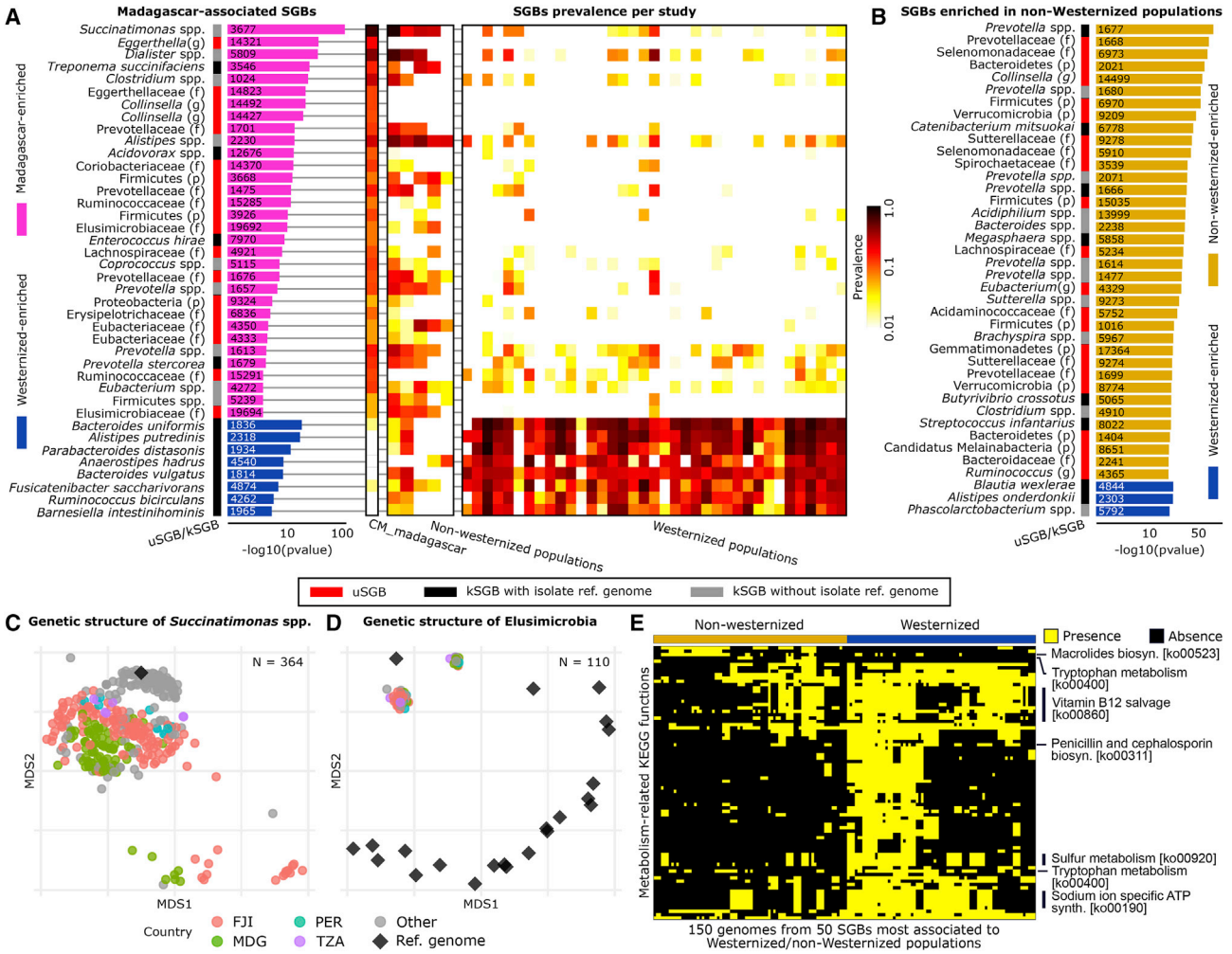


Figure 5. SGBs and Single Reconstructed Genomes Associated with Westernized and Non-Westernized Lifestyles

(A) 49 total large (>10 genomes) SGBs were significantly enriched (Fisher's test) in the set of 112 Madagascar gut metagenomes sequenced for this study, and 20 were significantly depleted (Fisher's test) relative to Western gut microbiomes (complete results in Table S6). Most Madagascar-enriched SGBs are uSGBs or contain only isolate sequences that were themselves assembled from other metagenomes in other studies.

(B) 232 total SGBs were differentially present with respect to the total set of non-Westernized populations, again with the 40 most significant—excluding those already reported in (A)—shown here (Fisher's test, complete results in Table S6).

(C) The intra-SGB genetic structure of *Succinatimonas* spp., the bacterium most associated with non-Westernized lifestyles (multidimensional scaling [MDS] on percentage nucleotide distances between genomes). The few genomes assembled from Westernized countries are tightly clustering together, while strains from non-Westernized populations are distinct and not well represented by the only available co-assembled (but not cultivated) strain.

(D) MDS of the two uSGBs (ID 19692 and ID 19694) enriched in the Madagascar cohort and available isolate genomes for the containing Elusimicrobia phylum (phylogeny in Figure S5A). The metagenomically assembled genomes in Elusimicrobia SGBs greatly diverge from the non-human-associated isolate genomes in the phylum.

(E) Significant differences in functional potential between the 25 SGBs most strongly associated with Westernized and non-Westernized populations. We report the differential KEGG pathways (Fisher's test Bonferroni-corrected $p < 0.05$, full list in Table S6) whose components are found in the set of representative genomes for the 50 species (only three genomes per SGB).

the same pathway (ko00400) differentially present in Westernized and non-Westernized communities (Figure 5E). UniRef50 annotations of all genomes highlighted many additional differences (82,563 with Bonferroni-corrected $p < 0.01$, Table S5), spanning also fimbrial functions and degradation of complex pectins enriched in the non-Westernized microbiomes. These associations of microbial functional potential with population capture a wide range of potential diet, metabolic, genetic, and exposure differences (De Filippo et al., 2010; Yatsunenko et al., 2012) and suggest that there are multiple ways in which the gut microbiome adapts to the diversity of human hosts.

DISCUSSION

This work expands the collection of microbial genomes associated with the human microbiome by more than doubling the