Genome-centric analysis of short and long read metagenomes reveals uncharacterized microbiome diversity in Southeast Asians

Jean-Sebastien Gounot[1]*, Chia Minghao[1]*, Denis Bertrand[1]*, Woei-Yuh Saw[2,3], Aarthi Ravikrishnan[1], Adrian Low[4], Yichen Ding[4], Ng Hui Qi Amanda[1], Linda Wei Lin Tan[5], Teo Yik-Ying[2,5,6#], Henning Seedorf[4,7#], Niranjan Nagarajan[1,8#]

[1]*Genome Institute of Singapore, Singapore 138672, Singapore*

[2]*Life Sciences Institute, National University of Singapore, Singapore 117456, Singapore*

[3] Baker Heart and Diabetes Institute, 75 Commercial Rd, Melbourne 3004, Victoria, Australia

[4]Temasek Life Sciences Laboratory, 1 Research Link, Singapore 117604, Singapore

[5]Saw Swee Hock School of Public Health, National University of Singapore, 12 Science Drive 2, Singapore 117549, Singapore

[6]Department of Statistics and Applied Probability, National University of Singapore, Singapore 117546, Singapore

[7]Department of Biological Sciences, National University of Singapore, Singapore 117558, Singapore

[8]*Yong Loo Lin School of Medicine, National University of Singapore, Singapore 117596, Singapore*

**Joint First Authors*

[#]Corresponding Authors

Lead Contact: nagarajann@gis.a-star.edu.sg

Abstract

Despite extensive efforts to address it, the vastness of uncharacterized 'dark matter' microbial genetic diversity can impact short-read sequencing based metagenomic studies. Population-specific biases in genomic reference databases can further compound this problem. Leveraging advances in long-read and Hi-C technologies, we deeply characterized 109 gut microbiomes from three ethnicities in Singapore to comprehensively reconstruct 4,497 medium and high-quality metagenome assembled genomes, 1,708 of which were missing in short-read only analysis and with >28× N50 improvement. Species-level clustering identified 70 (>10% of total) novel gut species out of 685, improved reference genomes for 363 species (53% of total), and discovered 3,413 strains that are unique to these populations. Among the top 10 most abundant gut bacteria in our study, one of the species and >80% of all strains were not represented in existing databases. Annotation of biosynthetic gene clusters (BGCs) uncovered more than 27,000 BGCs with a large fraction (36-88%) not represented in current databases, and with several unique clusters predicted to produce bacteriocins that could significantly alter microbiome community structure. These results reveal the significant uncharacterized gut microbial diversity in Southeast Asian populations and highlight the utility of hybrid metagenomic references for bioprospecting and disease-focused studies.

Introduction

While estimates for microbial diversity on Earth vary widely, studies suggest that there are nearly a million prokaryotic species of which only around 20,000 have been cultured[1,2]. The use of culture-free metagenomic techniques has therefore been key to unravel this 'dark matter' of genetic diversity on Earth. Microbial communities in a wide-range of biospheres have been explored, including terrestrial[3], aquatic[4] and extreme environments[5], as well as plant, animal and human-associated microbiomes[6]. Improvements in metagenomic assembly workflows[7–11] and computing resources have further enabled the assembly of these large datasets to construct metagenome-assembled genomes (MAGs) that serve to augment isolate-based reference genome databases[12,13]. Despite this, existing databases only represent approximately 48,000 species with genome sequences, and the accuracy and completeness of short-read based MAGs is frequently lower than isolate-based references[2].

Human gut metagenomes represent an area of intense scientific interest due to their association with various cancers, metabolic, immunological and neurological disease conditions[14,15]. Metagenome-wide association studies frequently rely on the completeness of reference genomes to correctly assign short reads to taxa, and link microbial genes and function to diseases[16]. In particular, existing studies suggest that there might be key population-specific differences in metagenomic associations with various diseases[17–19]. The availability of a large number of short-read metagenomic datasets (e.g. >20,000 for human gut in public repositories) has spurred the generation of MAG reference collections based on short-read assembly[13,20–22]. While these studies have added an impressive collection of genomes to existing databases, it is unclear yet if they are representative of the genetic diversity seen in gut metagenomes around the world. In addition, recent advances in sequencing assays (e.g. Hi-C[23], read cloud[24]), hybrid[25] and long-read metagenomic analysis[26] have sought to address the shortcomings of short-read metagenomics, and opened the possibility that long-read based MAGs can provide near-complete genomes rivaling isolate genomes in quality. As access to genome sequencing becomes democratized and gut metagenomes are explored in understudied populations, the strategy and value for establishing population-specific MAG references remains an open question.

Leveraging the availability of a multi-ethnic (Chinese, Malay and Indian) healthy adult cohort representing major Asian populations in Singapore, a city-state with high population density, we deeply characterized 109 gut metagenomes with state-of-the-art short read, long read and Hi-C technologies (*Singapore Platinum Metagenomes Project* – **SPMP**). The resulting datasets were assembled to produce high-quality references that significantly improve existing databases in assembly quality (>28× N50 improvement), helped identify 70 previously uncharacterized gut microbial species (>10% novel) and more than 3,400 strains in Southeast Asian populations, and uncovered thousands of novel BGCs that serve as a resource for bioprospecting. The ability to substantially augment existing databases through in-depth hybrid

metagenomic analysis highlights the value of this strategy, the importance of uncharacterized microbial diversity in Asia, and serves as a template for population-specific 'platinum' metagenome references for precision medicine programs around the world.

Results

Generation of a population-specific high quality gut microbial reference catalog

To explore the utility of various metagenomic strategies for generating a high-quality gut microbial reference database for a population, subjects from an existing multi-omics study in Singapore[27] were recruited to provide stool samples with informed consent (n=109; **Supplementary File 1**, **Methods**). Samples were collected using a kit designed for preserving anaerobes, DNA was extracted with a protocol optimized for high molecular weight, and shotgun sequencing was performed using short (Illumina, 2×151bp, average depth=9.4Gbp, **Supplementary File 2**) and long read (Oxford Nanopore Technologies - ONT, median N50=8.6kbp, average depth=5.8Gbp, **Supplementary File 2**) technologies, along with high-throughput chromosome conformation capture (Hi-C) analysis for a subset of samples (n=24; **Supplementary Figure 1**, **Supplementary File 2**, **Methods**). The distribution of taxa in both sequencing technologies (Illumina and ONT) were confirmed to be highly concordant (median correlation coefficient ρ=0.90), enabling joint analysis of both datasets (**Supplementary Figure 2**).

We next compared the commonly used short-read strategy for building MAG reference collections[13,20–22], with a recently proposed hybrid assembly strategy[25], for their utility in building a population-specific database (**Methods**). From a cost perspective, we noted that the hybrid strategy required <$150 in additional sequencing costs per sample (~100% increase in total cost) and marginal increase in cloud computing cost per sample (**Supplementary Note 1**). This in turn was observed to result in >61% increase in the number of genomes produced per sample (>15 additional MAGs; **Figure 1A**) with the hybrid strategy, with some samples yielding >80 genomes. Overall, 4,497 MAGs were obtained with hybrid assembly for 109 samples, versus 2,789 MAGs with short-reads alone (**Supplementary File 3**), with several abundant gut bacterial genera having enhanced representation within hybrid assemblies (e.g. *Bifidobacterium*, *Faecalibacterium* and *Blautia*; **Figure 1B**). This was observed to substantially improve read assignment to the reference genome database, ensuring that much fewer genomes were not detected, and with computed relative abundances being more consistent for hybrid assemblies versus short-read assemblies (**Figure 1C**). Overall, hybrid assemblies consistently improved the recovery of genomes across genera, with no significant bias to any specific genera, highlighting the versatility of this approach (**Supplementary Figure 3**).

Incorporation of long-read data in hybrid assemblies enabled marked improvements in assembly contiguity (>28×) as reported previously[25], with an average N50 of 339kbp (L50=12) with hybrid assembly relative to an N50 of 12kbp with short reads alone (**Figure 1D**). This was

also accompanied by a notably lower level of chimerism (<10% vs >20% with short-read assemblies) and similar annotated gene lengths as short-read assemblies (**Supplementary Figure 4**), suggesting that hybrid assemblies are robust to indel errors in long reads. Overall, this provided higher quality genomes based on MIMAG critera[28] after binning[10], where many hybrid MAGs had correctly reconstructed rRNA genes[29], and no such MAGs were obtained with short-read only assembly (**Figure 1E**, **Methods**). To assess if the quality of MAGs could be improved further, Hi-C data was used to assist in contig binning[30–35]. This was found to marginally increase the proportion of high-quality MAGs obtained, and double the proportion of near-complete genomes, with similar average assembly contiguity (**Supplementary Figure 5**, **Supplementary File 3**). As the per sample cost of Hi-C analysis is currently high (>$500), studies for generating population-specific references will need to consider this cost-benefit tradeoff.

Hybrid assembled genomes in SPMP were assigned taxonomy based on the Genome Taxonomy Database[2] (GTDB) and compared to existing reference genomes to assess their utility. SPMP genomes were found to provide notably improved references for most GTDB species, for both isolates (>6× increase in N50) as well as uncultivated organisms (>13×; **Figure 1F**). While the improvement in assembly is expected for uncultivated organisms that are primarily assembled using short-read metagenomics, the observed improvement for isolates (albeit smaller, Wilcoxon p-value=1.25×10^{-11}) is noteworthy as long-read sequencing is commonly used and the assembly problem is expected to be simpler. Overall, SPMP genomes provided high-quality references for 110 GTDB species, 46 of which have isolates, highlighting the value of a 'platinum' metagenomics approach for augmenting existing reference genome databases (**Figure 1G**).

Asian gut metagenomes harbor substantial uncharacterized gut microbial diversity

By encompassing three major Asian ethnicities (Chinese, Malay, Indian) in Singapore we anticipated that the SPMP would be a useful resource to explore Southeast Asian gut microbial diversity, and tested the idea of population-specific MAG reference catalogs (**Supplementary Figure 6**). Subsampling based rarefaction analysis with SPMP MAGs showed that with as few as a 100 subjects, >90% of the estimated recoverable (at the genomic level) gut microbial species diversity of the Singaporean population was represented in the SPMP catalog (**Figure 2A**, **Methods**). Similarly, with a reference genome collection that is 1/6th the size of a public gut microbial reference database[13] (UHGG; 18Gb vs 3Gb), SPMP can be used to identify more gut bacterial reads from an independent Singaporean study (manuscript under review; 92% vs 91%), and classify substantially more reads at the genome-level when database sizes are similar (81% vs 67%; **Supplementary Figure 7**). These results indicate that while the urban populations in Singapore have broadly similar representation of gut microbes, their genome sequences are still substantially distinct to impact mapping-based gut metagenome analyses.

To understand microbiome variability across ethnicities and its utility to discover new biological insights, we used multivariate regression analysis[36] to explore relationships between gut metagenome composition and demographic factors (e.g. sex, age, ethnicity). Interestingly, more than 60% of the taxonomic associations discovered (91 out of 133; FDR-adjusted p-value<0.05) were related to ethnicity, with 23 gender-specific and 19 age-based associations (**Supplementary File 4**). We then aggregated SPMP MAGs into species-level clusters (SLCs, 95% identity), annotating them with publicly available reference genome collections (**Supplementary Figure 8**, **Methods**) to identify 70 putative new species for which no genomes have been available previously, despite large-scale MAG generation efforts[2,13] (**Figure 2B**). Surprisingly, these putative new species represent >10% of the species-level clusters obtained (n=685) and are in addition to the 363 clusters that only have MAGs and no isolate genomes in existing databases (GTDB: https://gtdb.ecogenomic.org/, based on systematic analysis of curated genomes in RefSeq: https://www.ncbi.nlm.nih.gov/refseq/ and GenBank: https://www.ncbi.nlm.nih.gov/genbank/). More than 50% of the novel SLCs (38 out of 70) were only assembled with hybrid assembly and were missing in short-read assemblies. In addition, hybrid assemblies provided a >13× median N50 improvement overall, generating nearly all of the high-quality and near-complete genomes for the novel SLCs (19 out of 20), highlighting the utility of this strategy for capturing microbial diversity. In comparison to a recently published resource for under-represented East and South Asian populations[22] we found that most species were still novel (87%, 61/70) emphasizing the importance of generating population-specific references.

Among the novel SLCs, in addition to representatives in nearly all orders commonly containing gut microbes (e.g. Bacteroidales), we noted that 17 could be classified to the order Coriobacteriales while an additional 7 were assigned to Christensenellales, both of which are relatively understudied gut bacterial orders with high diversity in general and few isolates (**Supplementary Figure 9**). Additionally, three novel SLCs with high-quality MAGs represent the only available genomes for the corresponding genera (SLC637 – closest match *Phocaeicola*, <83% identity; SLC487 and SLC667 – closest match *Butyricicoccus*, <81% identity), while one of the novel SLCs is among the top 10 most abundant SLCs within the gut microbiomes of SPMP subjects (SLC612; **Supplementary Figure 10**). We noted that SLC612 is significantly more abundant in the gut microbiomes of Singaporean populations than in western subjects, potentially explaining why it was not assembled in previous large-scale studies, and emphasizing the need for population-specific references for even common gut bacteria (**Supplementary Figure 10**).

At the strain-level (99% identity), SPMP genomes were notably unique compared to >200,000 genomes in the UHGG database, with 3,413 novel strains out of 3,891 (87% novel, **Methods**). Among the top 20 most abundant gut bacterial species in SPMP, less than 20% of the strains were represented in UHGG, with only the keystone gut commensal *Bacteroides uniformis* having >40% of its strains being represented by genomes from other populations (**Figure 2C**). For

species that are extensively characterized due to their use as probiotics such as *Bifidobacterium adolescentis* and *Bifidobacterium longum*, we noted that while many strain genomes have been obtained from isolates (>30; **Supplementary Figure 11**), SPMP MAGs reveal an even greater uncharacterized diversity in the Singaporean population (>50 novel strains; **Figure 2C**, **Supplementary Figure 11**) that could be leveraged for probiotic discovery.

To explore the utility of the SPMP database for bioprospecting and discovering secondary metabolic pathways that may be important for gut microbiome structure and function, we combined comparative[37] and deep learning[38] based approaches for annotating biosynthetic gene clusters with high stringency filters (BGCs, **Methods**). In total, we identified 27,084 BGCs (DeepBGC: 23,175; antiSMASH: 3,909) that grouped into 16,055 gene cluster families by BiG-SCAPE[39] (GCFs; **Figure 2D**). More than 90% of the GCFs (15,134) did not display similarity to previously known BGCs in curated standard databases (antiSMASH and MIBiG) and were not found in annotations within an extensive collection of gut microbial reference genomes (HRGM**,** **Methods**), highlighting the value of using complementary algorithms for bioprospecting in new populations. We estimated that >85% of SPMP GCFs were not represented in curated databases, even when only a higher confidence set of predictions from antiSMASH was considered, while 49% of GCFs were novel even after taking into account more extensive HRGM antiSMASH annotations (**Supplementary Figure 12**, **13**).

While a significant fraction of GCFs were predicted to encode for saccharides (N=5,888, 37%), in line with their important functions in microbe-microbe and microbe-host interactions[40], many novel GCFs appear to encode diverse bioactive compounds such as ribosomally translated and post translationally modified peptides (RiPPs), polyketides and non-ribosomal peptides (NRPs) (**Figure 2D**), some of which may have antimicrobial function (**Supplementary Note 2**). In particular, a group of GCFs not represented in curated databases was predicted to synthesize a bacteriocin in a *Blautia* species, with 3 distinct gene configurations and genes encoding enzymes for peptide modification (radical SAM superfamily) and ABC transporter genes (GCF382/271/37, **Figure 2E**). Analyzing the structure of the microbial community in samples with and without the novel GCFs identified distinct networks, with presence of GCF382/271/37 associated with strong negative correlations between the *Blautia* species and multiple *Faecalibacterium* species including *Faecalibacterium prausnitzii* (**Figure 2F**, **Methods**). Together with the known role of *Faecalibacterium* species in gut health[41-42], these observations highlight the importance of comprehensively identifying secondary metabolic pathways for understanding gut metagenome function in human diseases.

Discussion

Despite the growing number of gut microbiome studies worldwide, including from remote populations in the Americas[43] and hunter-gatherer tribes in Africa[44], the gut microbial diversity of Asian populations remains understudied[45]. Singapore represents a microcosm of multiple

major Asian ethnic populations (Chinese, Malay and Indian) living in the shared environment of a modern metropolis. While there has been extensive study of gut metagenomes of ethnic Chinese individuals from China, fewer studies have involved individuals from Southeast Asia and India. The SPMP can thus represent an important reference for these populations, in addition to Singaporean studies. More broadly, we anticipate that the microbial diversity seen in SPMP might be similar to what would be observed in other major urban centers in Asia (e.g. New Delhi, Jakarta, Tokyo, Hong Kong), but is likely the 'tip of the iceberg' when considering rural and nomadic populations.

Various parameters are likely to define the appropriate strategy for a study similar to SPMP in other countries, including cost, targeted quality of reference genomes, ease of technology access, and availability of sufficient number of samples from a representative baseline cohort in the country. While we attempted to employ multiple different technologies for SPMP to get high-quality assemblies, we chose the middle-ground in terms of cost and accessibility as this is an important consideration for many countries. In particular, even higher-quality metagenomic assemblies are possible if HiFi reads from the Pacific Biosciences Sequel IIe system are available[46]. Also, the recent announcement of higher-quality reads from ONT could help improve assembly further and reduce costs[47]. Even as the sequencing landscape is constantly changing, the results from our study suggest that high-quality population-specific metagenomic references are already feasible with a modest-sized cohort and limited sequencing resources.

The advantages of having high-quality references for metagenomics are similar to what other areas of genetics and studies in model organisms have benefited from i.e. substantially reduced cost and effort in future studies by: (i) allowing the use of short reads or a single sequencing assay/technology, (ii) enabling increased sensitivity in identification of genomic features using reference-based approaches (e.g. taxonomic classifiers for metagenomics), (iii) ensuring that there are fewer 'dark matter' reads whose origin is unknown. We envisage that efforts such as SPMP will benefit the scientific community by spurring greater adoption of reference-based analyses in metagenome-wide association studies[48,49]. Additionally, as we noted in **Figures 1F** and **1G**, the quality of genomes that can be obtained using metagenomics is now comparable or better than what can be obtained from the sequencing of microbial isolates, especially with short reads. This can galvanize efforts to genetically map microbial ecosystems in diverse biospheres, further contributing to the references available to study human microbiomes, and understanding of strain sharing between humans and the environment. As sequencing costs, ease of use and accessibility of new technologies, and metagenomic assembly algorithms improve, we can expect that a majority of the high-quality microbial references that will be used in the future would be obtained through metagenomics, thus helping to bridge the

knowledge gap for the hundreds of thousands of microbial species that are estimated to be there on Earth.

The detection of 70 putative novel species in SPMP is perhaps not surprising given the unexplored microbial diversity and the limitations of current genetic databases. However, it is noteworthy that this is still a substantial fraction of the species detected in this study (>10%, **Figure 2B**), and while some of these species are not frequently detected across individuals, one of them was in the top 10 most abundant gut bacterial species, while others may still play a significant role in the biology of some individuals by being sporadically abundant (e.g. SLC665 which is among the top 20 most abundant species in 5% of subjects). Not surprisingly, at the strain-level an even larger fraction of the observed genetic diversity was novel, but what was notable was that this was true even for the more abundant and well-studied species in the gut microbiome (e.g. *Bacteroides uniformis* and *Bifidobacterium adolescentis*, **Figure 2C**). These observations highlight the overall value of such studies for discovering probiotic strains that could be leveraged for population health, with modest investments in metagenomic analysis cost (<$40,000), making it feasible for national microbiome projects around the world.

Finally, the identification of >23,000 BGCs in the SPMP database that were not represented in existing annotated databases (88% of total, **Figure 2D**) highlights that we are only scratching the surface in terms of harnessing microbial pathways and functions for synthetic biology and biotechnology applications. This was made possible by the high-contiguity of our hybrid assemblies (>28× N50 relative to short-read assemblies), and the characterization of distinct, underrepresented South-East Asian populations in SPMP harboring substantial novelty relative to curated BGC databases (>85%) and annotated reference genomes (49%, **Supplementary Figure 12**, **13**). The gut microbiome by virtue of being a dynamic, host-associated community with high diversity of microbes is a rich hunting ground for host-modulating, macro-nutrient catabolizing and micro-nutrient synthesizing functions[50,51]. In addition, homeostasis in the gut microbiome may be maintained by key members of the community through the selective expression of antimicrobial peptides[52] (AMPs), and correspondingly we identified hundreds of novel BGCs encoding putative bacteriocins, sactipeptides, lanthipeptides and lassopeptides that can now be further characterized (**Supplementary Note 2**). Notably, we found evidence that the presence of a BGC in a common *Blautia* species is associated with significant changes in overall gut microbiome community structure for SPMP subjects (**Figure 2F**). Together these results highlight the potential for novel AMPs discovered in SPMP to provide genetic templates for further optimization, and subsequent use to modulate the gut microbiome, or as new antimicrobials to target multi-drug resistant pathogens.

Figure Legends

Figure 1. Assembly strategy for high-quality microbiome references. (A) Boxplots showing the number of MAGs obtained across metagenomic datasets using short-read and hybrid assemblies (n=109). (B) Stacked barchart showing genus-specific breakdown of the number of MAGs obtained using short-read and hybrid assemblies (left) and boxplots for corresponding relative abundances of the genera (right). (C) Scatter-plot showing the relative abundance of *Bifidobacterium* genomes estimated using short-read or hybrid assemblies for a sample (y-axis) versus corresponding relative abundances obtained using the default Kraken2 database (x-axis). (D) Violin plots showing the distribution of a contiguity metric (N50 – largest contig size where >50% of the genome is in larger contigs) for short-read and hybrid assembly based MAGs. (E) Stacked barcharts showing the relative proportion of MAGs satisfying different MIMAG quality standards with short-read and hybrid assemblies of SPMP datasets. (F) Violin plots showing the relative improvement in contiguity (N50) obtained using hybrid assembly MAGs from SPMP relative to matched genomes in the GTDB database. (G) Barcharts showing the number of GTDB reference genomes which were improved from medium to high MIMAG quality using SPMP MAGs. Center lines in the boxplots represent median values, box limits represent upper and lower quartile values, whiskers represent 1.5 times the interquartile range above the upper quartile and below the lower quartile, and all data points are represented as dots in the figures.

Figure 2. Characterization of novel species, strains and gene families in SPMP genomes. (A) Collection curve analysis showing that the SPMP database covers a substantial fraction of the species level diversity in its MAGs. (B) Pie-chart showing the breakdown of species-level clusters in SPMP that have an *isolate* genome, only have MAGs (*uncultivated*) and are *novel* compared to genomes in public databases (UHGG, GTDB, SGB). (C) Stacked barcharts showing the number of SPMP strains that have an *isolate* genome, only have MAGs (*uncultivated*) and are *novel* compared to all UHGG genomes (>200,000, <99% ANI). The species shown are the top 20 in terms of median relative abundance in SPMP (most abundant on the left). (D) Stacked barcharts showing the number of BGCs (top) and GCFs (bottom) in different product classes that are present or absent in existing annotations comprising of the antiSMASH and MiBIG databases as well as antiSMASH annotations from HRGM. Inset piecharts show the overall breakdown. (E) Synteny plots showing the conservation of gene order and orientation (colored arrows, relatedness shown by vertical lines) for a novel GCF (GCF382) and related families. (F) Network diagrams depicting correlations between gut microbial species (nodes – species, edges – significant correlations) and overall microbiome structure in SPMP metagenomes when stratified based on presence or absence of GCF 382/271/37 (or missing the corresponding transporter gene) in a *Blautia* species (enlarged teal node, solid edges to correlated species, dashed edges between other nodes).

Methods

Subject recruitment

Subjects for this study were recruited based on recall from a community-based multi-ethnic prospective cohort[27] that is part of the Singapore Population Health Studies project (SPHS - formerly Singapore Consortium of Cohort Studies). Subjects in SPHS were recruited to participate in the National Health Survey, where subjects were selected at random using age- and gender-stratified sampling to obtain a representative sample set of residents in the country. At the point of recruitment in 2008, subjects did not have any pre-existing major health conditions (cardiovascular disease, mental illness, diabetes, stroke, renal failure, hypertension and cancer) based on self-reporting[27]. The ethnicity of each subject was confirmed verbally so that all four grandparents of the subject belonged to the same ethnic group. Informed consent was obtained from all participants and the associated protocols for this study were approved by the National University of Singapore Institutional Review Board (IRB reference number H-17-026).

Sample collection

Fecal samples were collected from healthy subjects using the BioCollector™ kit (The BioCollective, Colorado, USA). Samples were kept at -20°C until they were brought into an anaerobic chamber (atmosphere of N_2 (75%), CO_2 (20%) and H_2 (5%)). Fecal samples were homogenized and subsamples transferred into sterile 2 mL centrifuge tubes.

DNA extraction

Genomic DNA was extracted from fecal material (0.25 g wet weight) using the QIAamp Power Fecal Pro DNA kit (QIAGEN GmbH, Cat. No. 51804) and was quantified using Qubit dsDNA BR Assay Kit (Thermo Fisher Scientific, Cat. No. Q32853). Integrity of the extracted DNA was verified using 0.5% agarose gel electrophoresis.

Illumina library preparation and sequencing

Metagenomic libraries were prepared with a standard DNA input of 50ng across all samples, using NEBNext® Ultra™ II FS DNA Library Prep Kit for Illumina (New England Biolabs, Cat. No. E7805), according to the manufacturer's instructions. The reaction volumes were, however, scaled to a quarter of the recommended volumes for cost effectiveness. Barcoding and enrichment of libraries was carried out using NEBNext® Multiplex Oligos for Illumina® (96 Unique Dual Index Primer Pairs; New England Biolabs, Cat. No. E6440). Paired-end sequencing (2×151bp reads) was carried out on the Illumina HiSeq4K platform.

ONT library preparation and sequencing

Purity and integrity of DNA was assessed and ensured to fall within recommended ranges before library preparation. To preserve the integrity of DNA, the shearing step was omitted and DNA

was used directly for DNA repair and end-prep. Single-plex libraries were prepared using 1D sequencing kit (Oxford Nanopore Technologies, SQK-LSK108 or SQK-LSK109) according to the "1D Genomic DNA by ligation" protocol. For samples that were multiplexed (12-plex), the native barcoding kit (Oxford Nanopore Technologies, EXP-NBD103 or EXP-NBD104 and EXP-NBD114) was used and libraries were prepared according to the "Native barcoding genomic DNA" protocol. Both native barcode ligation and adapter ligation steps were extended to 30 min instead of 10 min. Single-plex samples were sequenced on either the MinION or GridION machine with either FLO-MIN106D or MIN106 revD flowcells. Multiplex samples were sequenced on the PromethION machine with FLO-PRO002 flowcells. Raw reads were basecalled with the latest version of the basecaller available at the point of sequencing (Guppy v3.0.4 to v3.2.6). Basecalled nanopore reads were demultiplexed and filtered for adapters with qcat (v1.1.0 https://github.com/nanoporetech/qcat).

Hi-C library preparation and sequencing

Hi-C libraries were generated using Phase Genomics ProxiMeta kit (version 3.0), based on the standard protocol. Briefly, 500 mg fecal material was crosslinked for 15 minutes at room temperature with end-over-end mixing in 1 mL of ProxiMeta crosslinking solution. Once crosslinking reaction was terminated, quenched fecal material was rinsed. Sample was resuspended and a low-speed spin was used to clear large debris. Chromatin was bound to SPRI beads and incubated for 1 hour with 150 µL of ProxiMeta fragmentation buffer and 11 µL of ProxiMeta fragmentation enzyme. Once washed, beads were resuspended with 100 µL of ProxiMeta Ligation Buffer supplemented with 5 µL of Proximity ligation enzyme and incubated for 4 hours. After reversing crosslinks, the free DNA was purified with SPRI and Hi-C junctions were bound to streptavidin beads and washed to remove unbound DNA. Washed beads were used to prepare paired-end deep sequencing libraries using ProxiMeta Library preparation reagents. Paired-end sequencing (2×151bp reads) was carried out on the Illumina HiSeq4K platform.

Sequence quality assessment

Illumina and ONT read statistics were generated with Fastq-Scan (v0.4.1, https://github.com/rpetit3/fastq-scan) and NanoStat[53] (v1.4.0), respectively. To assess taxonomic concordance, Illumina and ONT reads were classified with Kraken2[54] (v2.1.1, UHGG database[13]) and relative abundances were estimated with Bracken[55] (v2.6.1) at the species level (option -l R7) to compute Pearson correlation coefficients per sample.

Metagenomic assembly and binning

Illumina reads were assembled using MEGAHIT[8] (v1.04, default parameters) and hybrid metagenomic assemblies were generated with Illumina and ONT data using OPERA-MS[25] (v0.9.0, --polish). Contigs were binned with MetaBAT2[10] (v2.12.1, default parameters). Hi-C binning was

provided by Phase Genomics using its internal pipeline with MetaBAT results for hybrid assemblies as a starting point. Assembly bins were evaluated based on MIMAG standards[28], with contamination, completeness and N50 values determined with CheckM[56] (v1.04), and non-coding RNA annotations from barrnap (https://github.com/tseemann/barrnap) (v0.9) and tRNAscan-SE[57] (v2.0.5, default parameters). Assembly bins with contamination <10% and completeness >50% were designated as *medium quality* MAGs, those with contamination <5% and completeness >90% as *near complete* MAGs, and additionally near complete MAGs with complete 5S, 16S and 23S rRNA genes and at least 18 unique tRNA genes were classified as *high quality* MAGs. All other bins were classified as *low quality* and were removed from further analyses. In total, 4,497 medium quality, near complete and high quality MAGs were designated as being part of the SPMP database. Hybrid and short-reads assembly based MAGs were further assessed for chimerism with GUNC[58] (v1.0.4, detailed output). Coding sequence lengths obtained from Prodigal[59] (v2.6.3) calls were compared between the two datasets to assess the potential impact of long read indel errors on gene annotation. Concordant with prior work showing that hybrid metagenomic assemblies can have high base-pair accuracy[25], we also noted that SPMP MAGs independently assembled from distinct individual gut metagenomes could exhibit high average nucleotide identity (>99.99%, consistent with Q40 quality).

Annotation of MAGs with the Genome Taxonomy Database

The SPMP database was compared to the GTDB database[2] (release 95) using GTDBtk's[60] (v1.4.1) ani_rep command with default arguments, which leverages MASH[61] (v2.3) to provide pairwise genome-wide similarity values between all query MAGs and GTDB sequences. Only pairs with MASH distance $\leq$0.05 were retained and used to define the best match for each SPMP MAG based on minimum MASH distance. GTDB matches were classified based on their metadata as being *uncultivated* ("derived from environmental sample" or "derived from metagenome") or based on *isolate* strains. Both N50 values and MIMAG classifications were extracted from GTDB metadata. MAGs were placed into a phylogenetic tree using GTDB_TK (v1.4.1) with classify_wf (default options), based on pplacer_taxonomy values. To assess novelty in light of the latest human gut metagenome database, we further compared our MAGs to the 5,414 representative genomes from the Human Reference Gut Microbiome catalog (HRGM)[22] with a similar MASH analysis (**Supplementary File 5**).

Species and strain-level clustering

MAGs were clustered at the species (95%) and strain-level (99%) based on average nucleotide identity estimates (ANI; using MASH with sketch size of 10k and k-mer size of 21bp) with agglomerative clustering (sklearn v0.23.2, AgglomerativeClustering function, options: linkage="single", n_clusters=None, compute_full_tree=True, affinity="precomputed"). For each cluster, *representative* MAGs were defined using the highest eigen centrality value based on a weighted network graph produced by networkx (v2.5; eigenvector_centrality function). Strain-

level clustering was done jointly with all species-level matches from the UHGG database (v1.0, ANI threshold of 95%). Phylogenetic analysis at the strain-level was conducted using the biopython Phylo package[62], based on pairwise distances generated with FastANI[63] (v1.32). Phylogenetic trees were visualized using FigTree (tree.bio.ed.ac.uk/software/figtree).

Species assignment

Species-level clusters (SLCs) were assigned putative species name and types based on comparisons with multiple databases, including GTDB, Pasolli et al[64] (SGB) and Almeida et al[13] (UHGG). SLCs types were defined as, (i) isolate: if GTDB match to an isolate was found (mash distance $\leq$0.05), (ii) uncultivated: if a match to any database was found, but no isolates, (iii) novel: if no matches were found. SLCs were assigned putative species names based on a majority rule for MAGs in the cluster, with preference for GTDB ids (**Supplementary Figure 8**).

Species abundance and rarefaction analysis

Representative MAGs for SLCs were used to create a custom Kraken[65] (v2.1.1) database (https://github.com/DerrickWood/kraken2/wiki/Manual#custom-databases) and relative abundances for SLCs were estimated for each sample using Bracken[55] (v2.6.0, default parameters). Rarefaction analysis for estimating overall species diversity was done using the R package iNext[66] (v2.1.7, q=0, datatype="incidence_raw" and endpoint=300), based on converting SLC relative abundance values from Bracken into presence-absence values at a threshold of 0.05%.

Multivariate regression analysis

Genus-level abundances for each sample were provided as input for R package MaasLin2[36] (v1.4.0) along with sample metadata (age, sex and ethnicity), and significant associations were determined by combining 3 MaasLin2 runs with a compound Poisson linear model.

Biosynthetic gene cluster identification and clustering

Biosynthetic gene clusters (BGCs) in the SPMP database were identified using antiSMASH[67] (v5.1.2, --genefinding-tool prodigal-m --cb-general --cb-knownclusters --cb-subclusters --asf --pfam2go --smcog-trees) and DeepBGC[38] (v0.1.18, prodigal-meta-mode). BGCs with only one identified gene and with length <2kbp were removed for both sets of results. For antiSMASH this provided a set of 3,909 BGCs. DeepBGC results which overlapped with antiSMASH were removed if the genomic coordinates of both BGCs overlapped by ≥30% in either direction. DeepBGC candidates were further filtered for i) being categorized with a known product class and ii) containing at least one known biosynthetic pfam or TIGRFAM protein domain as defined by Cimermancic et al[68], providing an additional set of 23,175 BGCs.

All 27,084 BGCs (3,909 from antiSMASH + 23,175 from DeepBGC) were first categorized into different product classes: ribosomally synthesized and post-translationally modified peptides

(RiPPs), nonribosomal peptide synthetases (NRPs), polyketide synthases (PKS), saccharides and others based on the labels reported by each algorithm. We further unified the antiSMASH and DeepBGC product class labels to integrate both datasets (**Supplementary Table 1**). A fraction of mined BGCs were labeled as "hybrids" because antiSMASH or DeepBGC associated them with two different product classes e.g. "bacteriocin;T1PKS". The BGCs in each product class were grouped into gene cluster families (GCFs) by sequence similarity using BiG-SCAPE[39] (v1.01, --include_singletons --mix --no_classify --cutoffs 0.3). A total of 16,055 GCFs were defined by this approach and for each GCF we took the smallest BGC member as a representative of the family. Gene cluster diagrams of BGCs were created using Clinker[69].

BGCs in SPMP were classified as *novel* via a two-step approach. Firstly, BGC sequences were required to have <80% similarity to any existing sequence in the antiSMASH and MIBiG 2.0[70] databases using the clusterblast results from antiSMASH. Secondly, BGC annotations were compared to antiSMASH annotations from a comprehensive gut microbial genome collection (HRGM) using the standalone clusterblast software[71] (v 1.1.0), to identify SPMP matches based on a 80% similarity threshold, similar to the approach described in Gallagher et al[72].

Characterization of antimicrobial peptides and impact on microbiome structure

Antimicrobial activities of putative peptides encoded by novel RiPP BGCs in SPMP were predicted using an ensemble voting approach with four different antimicrobial peptide (AMP) prediction models: AMPscanner[73] (v2, convolutional neural network), AmpGram[74] (random forest model), AMPDiscover[75] (based on quantitative sequence activity models) and ABPDiscover (https://biocom-ampdiscover.cicese.mx/). Peptides predicted by antiSMASH in these RiPP BGCs were translated and all amino acid sequences with a length greater than 10 but lesser than 200 were used as inputs into these four models. Peptides were classified as AMPs if they received votes from both AMPscanner and AmpGram, and at least one vote from either AMPDiscover or ABPDiscover, and corresponding RiPP BGCs contained a transporter protein. The performance of this ensemble approach was evaluated using 78 known AMP sequences and 78 scrambled non-AMP sequences taken from the AmpGram benchmark dataset[74]. For our evaluation dataset, we identified and removed all sequences that were found in the training sets of AMPscanner, AmpGram, AMPDiscover and ABPDiscover using seqkit[76] (v0.11.0) and samtools faidx (v1.9). The percentage hydrophobicity and overall charge of selected peptide sequences was determined using the antimicrobial peptide calculator in the antimicrobial peptide database 3 (APD3; https://aps.unmc.edu/prediction).

To associate BGC presence/absence patterns with microbial community structure, correlation analysis (Fastspar[77] v1.0.0, parameters: --iterations 100 --exclude_iterations 20, p-values from 1000 bootstrap replicates and permutation testing) was done based on SLC abundance profiles across samples (species with medium abundance ≤0.1% filtered out). Correlations in the network were kept if they had an associated p-value <0.05.

Data and source code availability

Shotgun metagenomic sequencing data (Illumina and ONT) are available from the European Nucleotide Archive (ENA – https://www.ebi.ac.uk/ena/browser/home) under project accession number PRJEB49168. Source code for scripts used to analyze the data are available in a GitHub project at https://github.com/CSB5/SPMP.

References

1. Curtis, T. P., Sloan, W. T. & Scannell, J. W. Estimating prokaryotic diversity and its limits. *Proc. Natl. Acad. Sci.* **99**, 10494–10499 (2002).
2. Parks, D. H. *et al.* GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic Acids Res.* **1**, 13–14 (2021).
3. Nayfach, S. *et al.* A genomic catalog of Earth's microbiomes. *Nat. Biotechnol.* **39**, 499–509 (2021).
4. Tully, B. J., Graham, E. D. & Heidelberg, J. F. The reconstruction of 2,631 draft metagenome-assembled genomes from the global oceans. *Sci. Data* **5**, 1–8 (2018).
5. Tyson, G. W. *et al.* Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* **428**, 37–43 (2004).
6. Tierney, B. T. *et al.* The landscape of genetic content in the gut and oral human microbiome. *Cell Host Microbe* **26**, 283--295.e8 (2019).
7. Li, D., Liu, C. M., Luo, R., Sadakane, K. & Lam, T. W. MEGAHIT: An ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* **31**, 1674–1676 (2015).
8. Li, D. *et al.* MEGAHIT v1.0: A fast and scalable metagenome assembler driven by advanced methodologies and community practices. *Methods* **102**, 3–11 (2016).
9. Kolmogorov, M., Rayko, M., Yuan, J., Polevikov, E. & Pevzner, P. MetaFlye: Scalable long-read metagenome assembly using repeat graphs. *bioRxiv* 637637 (2019). doi:10.1101/637637
10. Kang, D. D. *et al.* MetaBAT 2: An adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ* **2019**, (2019).
11. Wu, Y. W., Simmons, B. A. & Singer, S. W. MaxBin 2.0: An automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics* **32**, 605–607 (2016).
12. Parks, D. H. *et al.* A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat. Biotechnol.* **36**, 996 (2018).
13. Almeida, A. *et al.* A unified catalog of 204,938 reference genomes from the human gut microbiome. *Nat. Biotechnol.* **39**, 105–114 (2021).
14. Kishikawa, T. *et al.* Metagenome-wide association study of gut microbiome revealed novel aetiology of rheumatoid arthritis in the Japanese population. *Ann. Rheum. Dis.* **79**, 103–111 (2020).
15. Zhu, F. *et al.* Metagenome-wide association of gut microbiome features for schizophrenia. *Nat. Commun.* **11**, (2020).
16. Wang, J. & Jia, H. Metagenome-wide association studies: Fine-mining the microbiome. *Nature Reviews Microbiology* **14**, 508–522 (2016).

17. Zhernakova, A. *et al.* Population-based metagenomics analysis reveals markers for gut microbiome composition and diversity. *Science* **352**, 565–569 (2016).

18. Falony, G. *et al.* Population-level analysis of gut microbiome variation. *Science* **352**, 560–564 (2016).

19. Breuninger, T. A. *et al.* Associations between habitual diet, metabolic disease, and the gut microbiota using latent Dirichlet allocation. *Microbiome* **9**, 1–18 (2021).

20. Zou, Y. *et al.* 1,520 reference genomes from cultivated human gut bacteria enable functional microbiome analyses. *Nat. Biotechnol.* **37**, 179–185 (2019).

21. Forster, S. C. *et al.* A human gut bacterial genome and culture collection for improved metagenomic analyses. *Nat. Biotechnol.* **37**, 186–192 (2019).

22. Kim, C. Y. *et al.* Human reference gut microbiome catalog including newly assembled genomes from under-represented Asian metagenomes. *Genome Med.* **13**, (2021).

23. Burton, J. N., Liachko, I., Dunham, M. J. & Shendure, J. Species-level deconvolution of metagenome assemblies with Hi-C-based contact probability maps. *G3 Genes, Genomes, Genet.* **4**, 1339–1346 (2014).

24. Bishara, A. *et al.* High-quality genome sequences of uncultured microbes by assembly of read clouds. *Nat. Biotechnol.* **36**, 1067–1080 (2018).

25. Bertrand, D. *et al.* Hybrid metagenomic assembly enables high-resolution analysis of resistance determinants and mobile elements in human microbiomes. *Nat. Biotechnol.* **37**, 937–944 (2019).

26. Kolmogorov, M., Yuan, J., Lin, Y. & Pevzner, P. A. Assembly of long, error-prone reads using repeat graphs. *Nat. Biotechnol.* **37**, 540–546 (2019).

27. Saw, W. Y. *et al.* Establishing multiple omics baselines for three Southeast Asian populations in the Singapore Integrative Omics Study. *Nat. Commun.* **8**, (2017).

28. Bowers, R. M. *et al.* Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nature Biotechnology* **35**, 725–731 (2017).

29. Parks, D. H. *et al.* Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nat. Microbiol.* **2**, 1533–1542 (2017).

30. Beitel, C. W. *et al.* Strain- and plasmid-level deconvolution of a synthetic metagenome by sequencing proximity ligation products. *PeerJ* **2014**, e415 (2014).

31. Burton, J. N., Liachko, I., Dunham, M. J. & Shendure, J. Species-level deconvolution of metagenome assemblies with Hi-C-based contact probability maps. *G3 Genes, Genomes, Genet.* **4**, 1339–1346 (2014).

32. Du, Y. & Sun, F. HiCBin: Binning metagenomic contigs and recovering metagenome-assembled genomes using Hi-C contact maps. *bioRxiv* 2021.03.22.436521 (2021). doi:10.1101/2021.03.22.436521

33. Baudry, L., Foutel-Rodier, T., Thierry, A., Koszul, R. & Marbouty, M. MetaTor: A

computational pipeline to recover high-quality metagenomic bins from mammalian gut proximity-ligation (Meta3C) libraries. *Front. Genet.* **10**, 753 (2019).

34. Press, M. *et al.* Hi-C deconvolution of a human gut microbiome yields high-quality draft genomes and reveals plasmid-genome interactions. *bioRxiv* 198713 (2017). doi:10.1101/198713

35. Demaere, M. Z. & Darling, A. E. Bin3C: Exploiting Hi-C sequencing data to accurately resolve metagenome-assembled genomes. *Genome Biol.* **20**, (2019).

36. Mallick, H. *et al.* Multivariable association discovery in population-scale meta-omics studies. *bioRxiv* 2021.01.20.427420 (2021). doi:10.1101/2021.01.20.427420

37. Medema, M. H. *et al.* AntiSMASH: Rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. *Nucleic Acids Res.* **39**, W339 (2011).

38. Hannigan, G. D. *et al.* A deep learning genome-mining strategy for biosynthetic gene cluster prediction. *Nucleic Acids Res.* **47**, e110 (2019).

39. Navarro-Muñoz, J. C. *et al.* A computational framework to explore large-scale biosynthetic diversity. *Nat. Chem. Biol.* **16**, 60–68 (2020).

40. Brouns, F. Saccharide characteristics and their potential health effects in perspective. *Frontiers in Nutrition* **7**, 75 (2020).

41. Lopez-Siles, M., Duncan, S. H., Garcia-Gil, L. J. & Martinez-Medina, M. Faecalibacterium prausnitzii: From microbiology to diagnostics and prognostics. *ISME Journal* **11**, 841–852 (2017).

42. Yao, Q. *et al.* Potential of fecal microbiota for detection and postoperative surveillance of colorectal cancer. *BMC Microbiol.* **21**, (2021).

43. Clemente, J. C. *et al.* The microbiome of uncontacted Amerindians. *Sci. Adv.* **1**, (2015).

44. Schnorr, S. L. *et al.* Gut microbiome of the Hadza hunter-gatherers. *Nat. Commun.* **5**, (2014).

45. Dehingia, M. *et al.* Gut bacterial diversity of the tribes of India and comparison with the worldwide data. *Sci. Rep.* **5**, 18563 (2015).

46. Bickhart, M. *et al.* Generation of lineage-resolved complete metagenome-assembled genomes by precision phasing. *bioRxiv* 2021.05.04.442591 (2021). doi:10.1101/2021.05.04.442591

47. Sereika, M. *et al.* Oxford Nanopore R10.4 long-read sequencing enables near-perfect bacterial genomes from pure cultures and metagenomes without short-read or reference polishing. *bioRxiv* 2021.10.27.466057 (2021). doi:10.1101/2021.10.27.466057

48. Sanna, S. *et al.* Causal relationships among the gut microbiome, short-chain fatty acids and metabolic diseases. *Nature Genetics* **51**, 600–605 (2019).

49. Lloyd-Price, J. *et al.* Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature* **569**, 655–662 (2019).

50. Milshteyn, A., Colosimo, D. A. & Brady, S. F. Accessing Bioactive Natural Products from the Human Microbiome. *Cell Host and Microbe* **23**, 725–736 (2018).

51. Wilson, M. R., Zha, L. & Balskus, E. P. Natural product discovery from the human microbiome. *Journal of Biological Chemistry* **292**, 8546–8552 (2017).

52. Ostaff, M. J., Stange, E. F. & Wehkamp, J. Antimicrobial peptides and gut microbiota in homeostasis and pathology. *EMBO Mol. Med.* **5**, 1465–1483 (2013).

53. De Coster, W., D'Hert, S., Schultz, D. T., Cruts, M. & Van Broeckhoven, C. NanoPack: visualizing and processing long-read sequencing data. *Bioinformatics* 237180 (2018). doi:10.1093/bioinformatics/bty149

54. Wood, D. E., Lu, J. & Langmead, B. Improved metagenomic analysis with Kraken 2. *Genome Biol.* **20**, 1–13 (2019).

55. Lu, J., Breitwieser, F. P., Thielen, P. & Salzberg, S. L. Bracken: estimating species abundance in metagenomics data. *PeerJ Comput. Sci.* **3**, e104 (2017).

56. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: Assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **25**, 1043–1055 (2015).

57. Lowe, T. M. & Chan, P. P. tRNAscan-SE On-line: integrating search and context for analysis of transfer RNA genes. *Nucleic Acids Res.* **44**, W54--7 (2016).

58. Orakov, A. *et al.* GUNC: detection of chimerism and contamination in prokaryotic genomes. *Genome Biol.* **22**, (2021).

59. Hyatt, D. *et al.* Prodigal: Prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 1–11 (2010).

60. Chaumeil, P. A., Mussig, A. J., Hugenholtz, P. & Parks, D. H. GTDB-Tk: A toolkit to classify genomes with the genome taxonomy database. *Bioinformatics* **36**, 1925–1927 (2020).

61. Ondov, B. D. *et al.* Mash: Fast genome and metagenome distance estimation using MinHash. *Genome Biol.* **17**, 132 (2016).

62. Cock, P. J. A. *et al.* Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422–1423 (2009).

63. Jain, C., Rodriguez-R, L. M., Phillippy, A. M., Konstantinidis, K. T. & Aluru, S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat. Commun.* **9**, 5114 (2018).

64. Pasolli, E. *et al.* Extensive unexplored human microbiome diversity revealed by over 150,000 genomes from metagenomes spanning age, geography, and lifestyle. *Cell* **176**, 649--662.e20 (2019).

65. Wood, D. E. & Salzberg, S. L. Kraken: Ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.* **15**, R46 (2014).

66. Hsieh, T. C., Ma, K. H. & Chao, A. iNEXT: an R package for rarefaction and extrapolation of species diversity (Hill numbers). *Methods Ecol. Evol.* **7**, 1451–1456 (2016).

67. Blin, K. *et al.* AntiSMASH 5.0: Updates to the secondary metabolite genome mining pipeline. *Nucleic Acids Res.* **47**, W81–W87 (2019).

68. Cimermancic, P. *et al.* Insights into secondary metabolism from a global analysis of prokaryotic biosynthetic gene clusters. *Cell* **158**, 412–421 (2014).

69. Gilchrist, C. L. M. & Chooi, Y.-H. clinker & clustermap.js: automatic generation of gene cluster comparison figures. *Bioinformatics* **37**, 2473–2475 (2021).

70. Kautsar, S. A. *et al.* MIBiG 2.0: A repository for biosynthetic gene clusters of known function. *Nucleic Acids Res.* **48**, D454--D458 (2020).

71. Medema, M. H., Takano, E. & Breitling, R. Detecting Sequence Homology at the Gene Cluster Level with MultiGeneBlast. *Mol. Biol. Evol.* **30**, 1218 (2013).

72. Gallagher, K. A. & Jensen, P. R. Genomic insights into the evolution of hybrid isoprenoid biosynthetic gene clusters in the MAR4 marine streptomycete clade. *BMC Genomics* **16**, 1–13 (2015).

73. Veltri, D., Kamath, U. & Shehu, A. Deep learning improves antimicrobial peptide recognition. *Bioinformatics* **34**, 2740–2747 (2018).

74. Burdukiewicz Michałand Sidorczuk, K. *et al.* Proteomic screening for prediction and design of antimicrobial peptides with ampgram. *Int. J. Mol. Sci.* **21**, 1–13 (2020).

75. Pinacho-Castellanos, S. A., García-Jacas, C. R., Gilson, M. K. & Brizuela, C. A. Alignment-free antimicrobial peptide predictors: Improving performance by a thorough analysis of the largest available data set. *J. Chem. Inf. Model.* **61**, 3141–3157 (2021).

76. Shen, W., Le, S., Li, Y. & Hu, F. SeqKit: A cross-platform and ultrafast toolkit for FASTA/Q file manipulation. *PLoS One* **11**, (2016).

77. Watts, S. C., Ritchie, S. C., Inouye, M. & Holt, K. E. FastSpar: Rapid and scalable correlation estimation for compositional data. *Bioinformatics* **35**, 1064–1066 (2019).

Figure 1

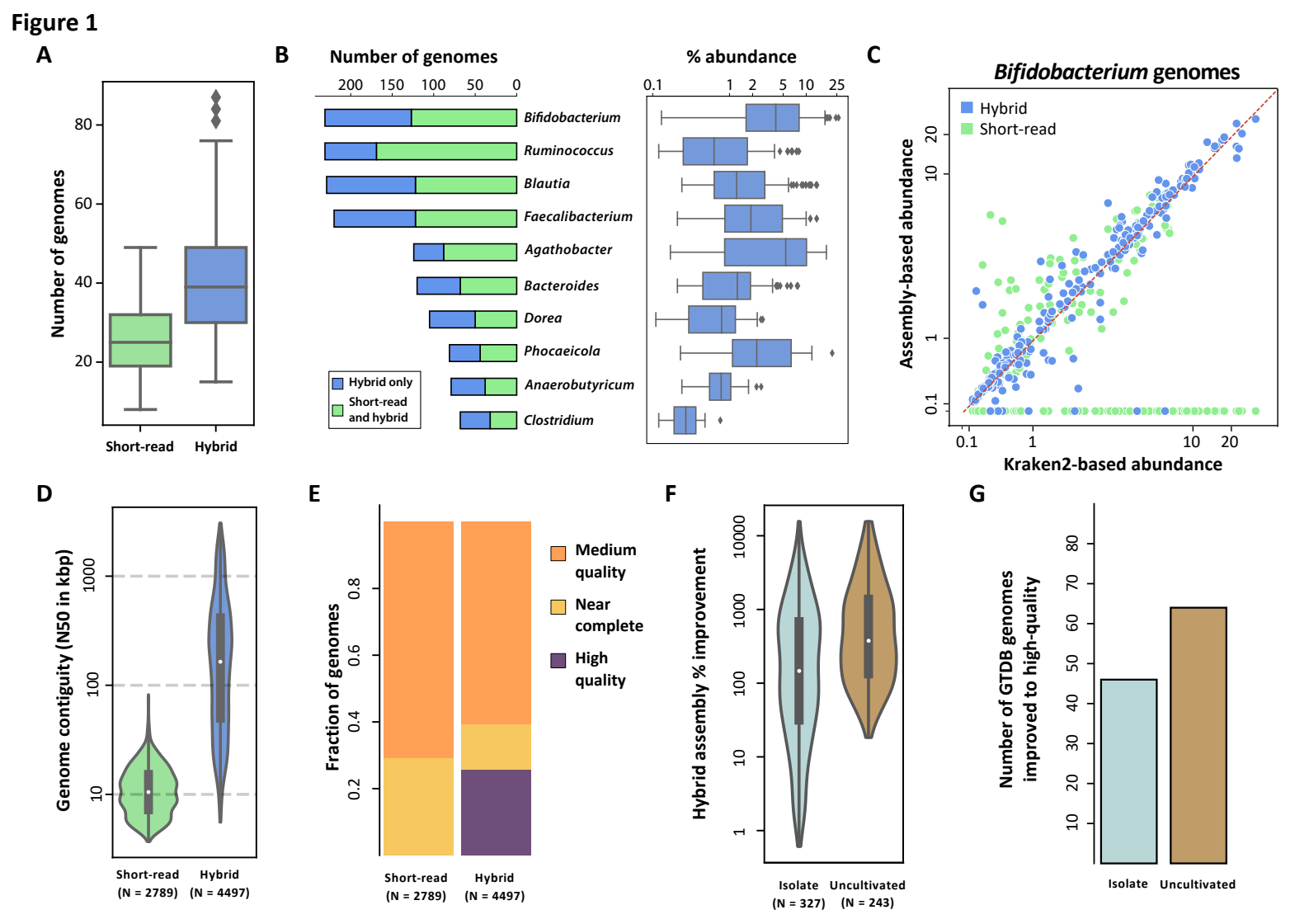
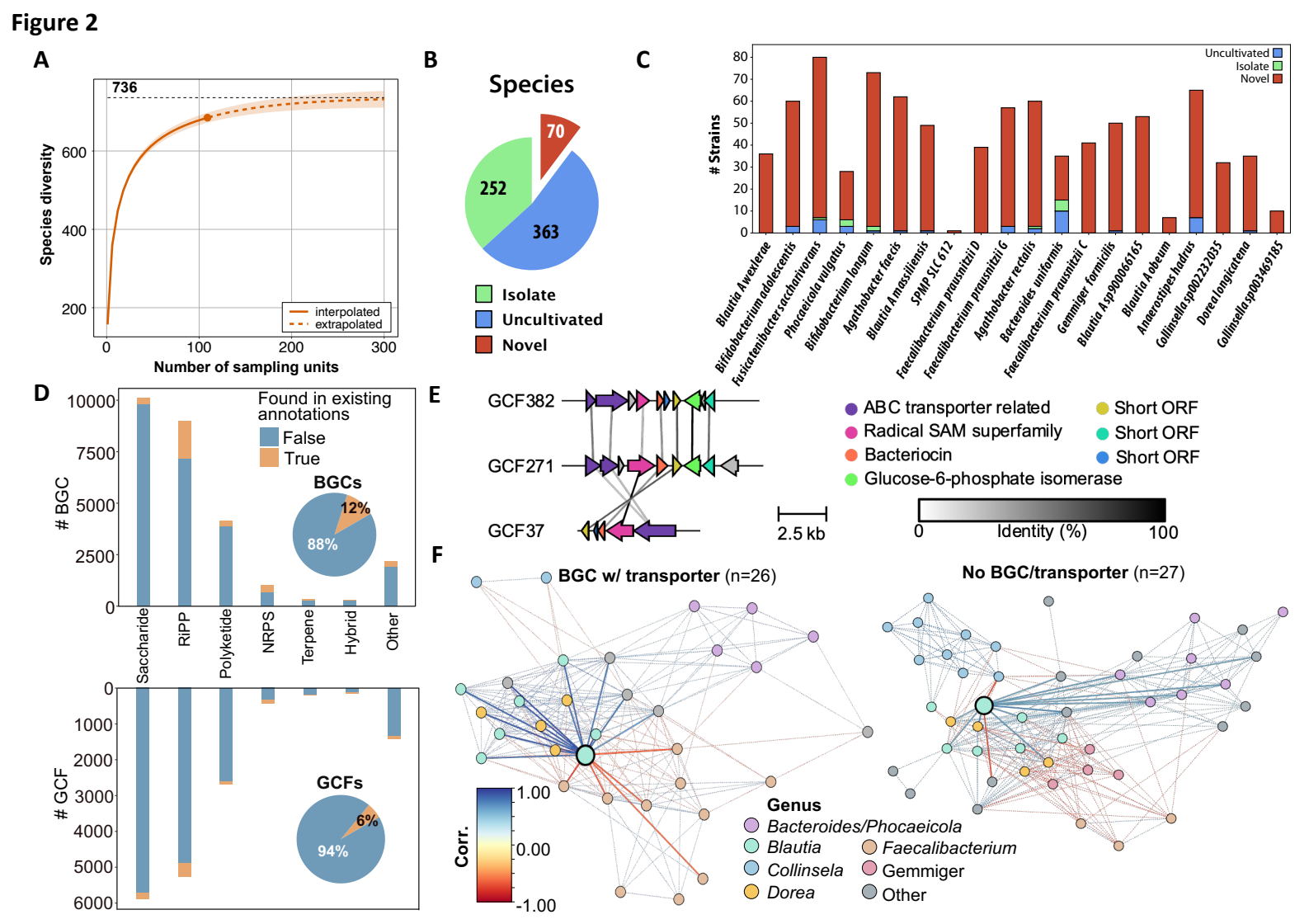

A
Number of genomes
80
60
40
20
Short-read
Hybrid
B
Number of genomes
200 150 100 50 0
% abundance
0.1 1 2 5 10 25
Bifidobacterium
Ruminococcus
Blautia
Faecalibacterium
Agathobacter
Bacteroides
Dorea
Phocaeicola
Anaerobutyricum
Clostridium
Hybrid only
Short-read and hybrid
C
Bifidobacterium genomes
Hybrid
Short-read
Assembly-based abundance
20
10
1
0.1
0.1
1
10
20
Kraken2-based abundance
D
Genome contiguity (N50 in kbp)
1000
100
10
Short-read (N = 2789)
Hybrid (N = 4497)
E
Fraction of genomes
0.8
0.6
0.4
0.2
Medium quality
Near complete
High quality
Short-read (N = 2789)
Hybrid (N = 4497)
F
Hybrid assembly % improvement
10000
1000
100
10
1
Isolate (N = 327)
Uncultivated (N = 243)
G
Number of GTDB genomes improved to high-quality
80
70
60
50
40
30
20
10
Isolate
Uncultivated

Figure 2


A
736
Species diversity
600
400
200
0
100
200
300
Number of sampling units
interpolated
extrapolated
B
Species
70
252
363
Isolate
Uncultivated
Novel
C
Uncultivated
Isolate
Novel
Strains
80
70
60
50
40
30
20
10
0
Blautia Awexlerae
Bifidobacterium adolescentis
Fusicatenibacter saccharivorans
Phocaeicola vulgatus
Bifidobacterium longum
Agathobacter faecis
Blautia A massiliensis
SPMP SLC 612
Faecalibacterium prausnitzii D
Faecalibacterium prausnitzii G
Agathobacter rectalis
Bacteroides uniformis
Faecalibacterium prausnitzii C
Gemmiger formicilis
Blautia A sp900066165
Blautia A obeum
Anaerostipes hadrus
Collinsella sp002232035
Dorea longicatena
Collinsella sp003469185
D
Found in existing annotations
False
True
BGC
10000
7500
5000
2500
0
BGCs
12%
88%
Saccharide
RiPP
Polyketide
NRPS
Terpene
Hybrid
Other
GCF
0
1000
2000
3000
4000
5000
6000
GCFs
6%
94%
E
GCF382
GCF271
GCF37
2.5 kb
ABC transporter related
Radical SAM superfamily
Bacteriocin
Glucose-6-phosphate isomerase
Short ORF
Short ORF
Short ORF
0
Identity (%)
100
F
BGC w/ transporter (n=26)
No BGC/transporter (n=27)
Corr.
1.00
0.00
-1.00
Genus
Bacteroides/Phocaeicola
Blautia
Collinsela
Dorea
Faecalibacterium
Gemmiger
Other