scientific data

Check for updates

OPEN

DATA DESCRIPTOR

A compilation of fecal microbiome shotgun metagenomics from hematopoietic cell transplantation patients

Jinyuan Yan[1✉], Chen Liao[1], Bradford P. Taylor[2], Emily Fontana[3], Luigi A. Amoretti[3], Roberta J. Wright[3], Eric R. Littmann[4,11], Anqi Dai[5], Nicholas Waters[5], Jonathan U. Peled[5,6], Ying Taur[3], Miguel-Angel Perales[5,6], Benjamin A. Siranosian[7], Ami S. Bhatt[7,8,9], Marcel R. M. van den Brink[5,6], Eric G. Pamer[4], Jonas Schluter[10] & Joao B. Xavier[1✉]

Hospitalized patients receiving hematopoietic cell transplants provide a unique opportunity to study the human gut microbiome. We previously compiled a large-scale longitudinal dataset of fecal microbiota and associated metadata, but we had limited that analysis to taxonomic composition of bacteria from 16S rRNA gene sequencing. Here we augment those data with shotgun metagenomics. The compilation amounts to a nested subset of 395 samples compiled from different studies at Memorial Sloan Kettering. Shotgun metagenomics describes the microbiome at the functional level, particularly in antimicrobial resistances and virulence factors. We provide accession numbers that link each sample to the paired-end sequencing files deposited in a public repository, which can be directly accessed by the online services of PATRIC to be analyzed without the users having to download or transfer the files. Then, we show how shotgun sequencing enables the assembly of genomes from metagenomic data. The new data, combined with the metadata published previously, enables new functional studies of the microbiomes of patients with cancer receiving bone marrow transplantation.

Background & Summary

The composition of gut microbiome changes in response to mild perturbations such as changes in diet[1] and strong perturbations such as chemotherapy[2] and antibiotics[3] that can deplete the majority of the microbes and impact microbiome function[3]. Over the past decades, the microbiome field has sought to characterize compositional changes to perturbations and understand how those changes impact human health[4]. Cross-sectional or longitudinal multi-omics data yielded valuable insights into the population dynamics of gut microbes, their ecological interactions and metabolic functions, and the molecular mechanisms of host-microbe crosstalk[5,6]. Data from patients hospitalized to receive allogeneic hematopoietic cell transplantation (HCT) provide a unique chance to study the gut microbiome in extremely perturbed conditions[7–9]. These perturbations caused by the treatment occur in a planned, scheduled fashion as patients stay in the hospital for several weeks, which enables collecting samples and clinical metadata. The patients receive many drugs including antibiotics that impact the composition and function of the gut microbiome[10,11]. The data also allow us to study how the microbiome

[1]Program for Computational and Systems Biology, Memorial Sloan-Kettering Cancer Center, New York, NY, USA. [2]Center for Communicable Disease Dynamics, Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA, USA. [3]Infectious Disease Service, Department of Medicine, and Immunology Program, Sloan Kettering Institute, New York, NY, USA. [4]Duchossois Family Institute, University of Chicago, Chicago, IL, USA. [5]Adult Bone Marrow Transplantation Service, Department of Medicine, Memorial Sloan Kettering Cancer Center, New York, NY, USA. [6]Weill Cornell Medical College, New York, NY, USA. [7]Department of Genetics, Stanford University, Stanford, CA, USA. [8]Department of Medicine, Division of Hematology, Stanford University, Stanford, CA, USA. [9]Department of Medicine, Division of Blood and Marrow Transplantation and Cellular Therapy, Stanford University School of Medicine, Stanford, CA, USA. [10]Institute for Computational Medicine, Department of Microbiology, New York University, New York, NY, USA. [11]Deceased: Eric R. Littmann. ✉e-mail: yanj2@mskcc.org; xavierj@mskcc.org

composition feeds back on the state of its living host, and address some basic science questions such as how the microbiome influences the dynamics of the human immune system[12].

We previously published the first data descriptor of our institutional microbiome dataset of HCT patients (>10,000 samples from >1,000 patients), where we compiled patients' gut microbiota compositions based on 16S rRNA gene sequencing of fecal samples and its associated metadata[13]. Subsets of this comprehensive dataset were analyzed in a number of publications[7,12,14–22]. Metagenomic shotgun sequencing is more expensive but has advantages compared to 16S rRNA gene sequencing[23]: it not only reveals the composition of the gut microbiome but also the functions encoded by the genes in the microbiome[24,25]. Bioinformatic tools that analyze shotgun sequencing data for different purposes—taxonomic classification of microbial composition[26], gene abundance prediction of specialty genes such as antibiotic resistance[27,28] and virulence factors[28], genome identification of strain-level or species-level metagenome-assembled genomes (MAGs)[29,30] and metabolic model reconstruction that translate the DNA sequences to biochemical reactions[31–33]—are now readily available. Some of these tools even work directly with the accession numbers of the sequencing data deposited in public repositories, which greatly facilitates analysis.

Here we compile 395 human fecal samples that were analyzed by metagenomic shotgun sequencing. This compilation is a nested subset of samples we published previously but had analyzed only by 16S rRNA amplicon sequencing[13]. Here we present examples of functional analyses enabled by shotgun sequencing: metagenome functions such as virulence factors and antibiotic resistance and the assembly of genomes from metagenomic data. We first conduct a data validation where we check the data for quality by addressing specific questions: Do the compositions inferred from metagenomic and 16S sequencing data agree? How well does metagenomic sequencing capture antibiotic resistance genes? Can the metagenomic data recapitulate the genomic difference of bacterial pathogens? We display the 395 shotgun samples on a t-SNE map of the >10,000 samples of 16S amplicon sequencing[13]. We then investigate correlations between the consistency of stool samples and the read counts of shotgun samples, and we check the correlation of composition between 16S amplicon sequencing and shotgun metagenomes. We validate the ability to detect antibiotic resistance genes using an orthogonal detection of the *vanA* gene for vancomycin resistance using a PCR test. We used the available tools from PATRIC, a publicly accessible database and tool repository for bacterial genome analysis, to do compositional analysis (kranken2), virulence gene (VFDB) and antibiotic resistant gene (CARD) identification. We assembled metagenomically assembled genomes (MAGs) from shotgun reads and compared them with genomes sequenced from isolates of *Enterococcus faecium* obtained from the same samples[34]. We provide Matlab code to compile the output of these metagenomic analysis tools in a Github repository https://github.com/Jinyuan1998/scientific_data_metagenome_shotgun. The shotgun sequencing data together with the metadata provided earlier[13] enables longitudinal studies of the gut microbiome in patients hospitalized to receive bone marrow transplantation.

Methods

Ethics process of sample collection. Sample collection from patients and analysis of the biospecimens were approved by the Memorial Sloan Kettering Cancer Center Institutional Review Board. Signed informed consent for specimen collection were provided by all participants. All clinical metadata were formatted following the same rule as a previous publication[13]: The PatientID is a non-identifiable patient number that can be used to link clinical metadata to microbiota sample data, and all event dates of any patient were made to be relative to a patient-specific, deidentified reference date (see column 'Timepoint'). The secret reference dates will not be disclosed. We also provided sample collection dates relative to the date of nearest HCT of any patient (see column 'DayRelativeToNearestHCT').

Choice of samples to include in this study. The 395 samples in this compilation were chosen for shotgun sequencing analysis for diverse reasons: some samples were singled out for their importance in testing specific functional analyses, which others where sequenced because a preliminary analysis of the 16S taxonomic representations showed intriguing patterns. We included all the samples that we had shotgun sequenced until the submission date rather than excluding any samples. Therefore, rather than a single criterion of choice the data compiled gather a nested subset that is heterogenous but provides reasonable coverage of the microbiota states experienced by these patients, as determined by the analysis shown in Fig. 1.

Library preparation, shotgun sequencing and human genome decontamination. We compiled 395 of the >10,000 stool samples acquired from allo-HCT patients[13], extracted the genomic DNA and sequenced on the Illumina HiSeq platform as described previously[12,15]. We removed normal optical duplicates in paired FASTQ files using the clumpify.sh tool from the BBMap package, producing a pair of read files without duplicates. Using the bbduk.sh script in the BBMap package, we trimmed the right and left side of a read in a pair to Q10 using the Phred algorithm. A pair of reads was dropped if any one of them had a length ≤51 nucleotides after trimming[35]. We trimmed 3′-end adapters using a kmer of length 31, and a shorter kmer of 9 at the other end of the read. One mismatch was allowed in this process, and we allowed adapter trimming based on pair overlap detection (which does not require known adapter sequences) using the 'tbo' parameter. We used the 'tpe' parameter to trim the pair of reads to the same length. We removed human contamination using Kneaddata employing BMTagger. The BMTagger database was built with human genome assembly GRCh38. The paired end read files were uploaded into the Short Read Archive (SRA) of the National Center for Biotechnology Information (NCBI)[36,37].

Taxonomy classification and specific gene mapping for metagenomic reads. We used the services provided by the Pathosystems Resource Integration Center (PATRIC)[38]. PATRIC can take input as the SRA accession number of each sample and output the microbiome composition in taxa, as well as genes encoding

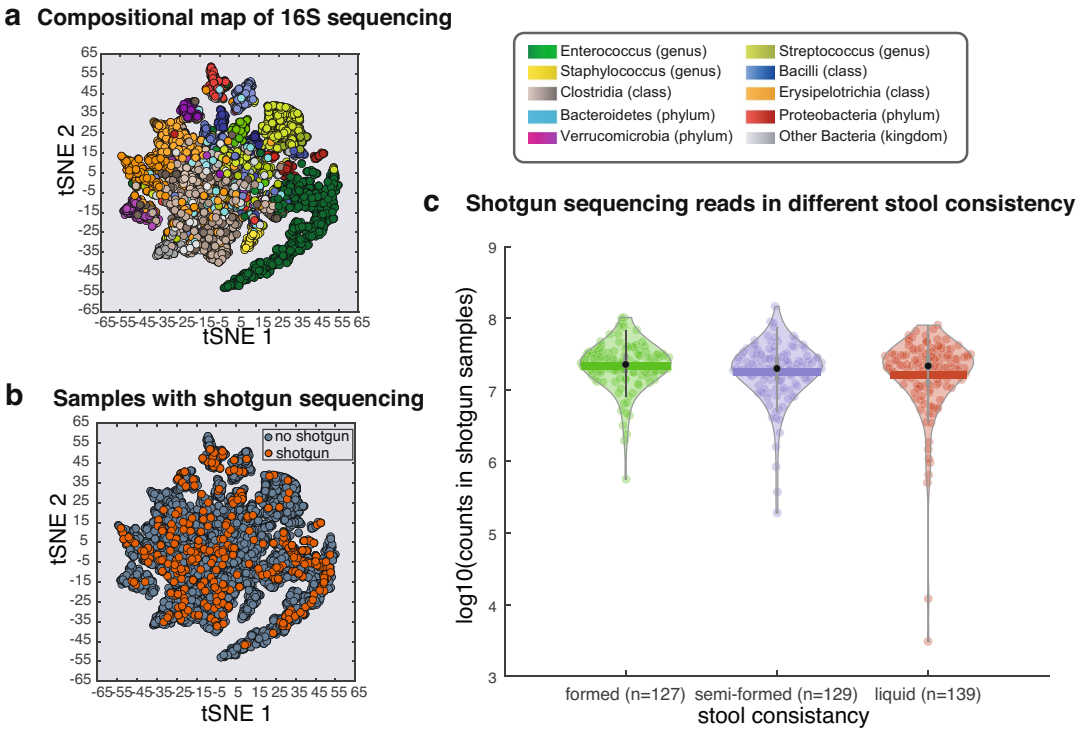

Fig. 1 The metagenomic samples cover the majority of microbiome compositional states observed in fecal samples from allo-HCT patients. (**a**) The t-SNE plot built using the taxonomic composition obtained by 16S amplicon sequencing of >10,000 samples from >1,000 unique patients;[13] the different colors indicate the most abundant taxon in each sample. (**b**) Location of nested subset of 395 samples from 49 unique patients with shotgun sequencing is broadly distributed across the entire map. (**c**) The sequencing depth of shotgun sequenced samples varies between 10^6 reads to 10^8 reads, with outliers in the liquid samples whose microbiome may yield different sizes of libraries.

virulence factors and antibiotic resistances. It uses the algorithm Kraken 2[26] for taxonomic classification, and the algorithm KMA[39] to align the metagenomic reads to non-redundant databases. The virulence factor composition analysis is based on the Virulence Factor Database[28] and the antibiotic resistance composition is based on the Comprehensive Antibiotic Resistance Database (CARD)[27]. The taxonomy, virulence factor and antibiotic resistance table for each of the 395 samples are provided as text tables.

Comparison between 16S data and shotgun metagenome data. Our analysis of shotgun vs 16S taxonomic representations uses the abundance tables produced by the PATRIC tool—explained above—and compares that output with the taxonomic abundances that we had obtained earlier for the 16S data[13]. That comparison required normalizing the taxonomic abundances. In Table 1 we list the details used for this normalizing and criteria for inclusion in the 16S data vs. shotgun metagenome data comparison. We focus here on details of the shotgun analysis, since the 16S analysis has already been published and its details explained in our previous publication[13]. For more details on the shotgun analysis the interested reader may obtain the Matlab code used for this comparison in the Github repository https://github.com/Jinyuan1998/scientific_data_metagenome_shotgun.

Genome assembly. We adapted a recently published pipeline to assemble the genomes of bacteria from shotgun sequenced samples[40]. Briefly, the pipeline first assembled contigs using metaSPAdes[41]. Then, it binned the contigs into MAGsusing three different methods: Metabat2[29] CONCOCT[30] and Maxbin2[42]. The results were then aggregated using DASTool which implements a dereplication, aggregation and scoring strategy[43] to produce the strain-level genomes.

Computer code for reproducible analysis. To make our analysis fully reproducible we provide the computer code for the analyses listed below in the GitHub repository https://github.com/Jinyuan1998/scientific_data_metagenome_shotgun. The repository included instructions to run the software, including parameters and databases used for preprocessing and the code is commented to denote the relevant analysis parameters.

Sequencing batches. The sample collection, preparation and preprocessing steps were kept constant between batches. It is important, however, that future analyses will be able to confirm there was no batch effect, or if there was one, what effects it had on the results. The samples.csv data table includes a column indicating the pool and the run IDs for 16S-sequenced samples and a batch ID for shotgun-sequenced samples.

Analysis step	Input	Output	Procedure
Normalization of shotgun taxa	Table of reads aligned to taxa produced by the Kraken2 pipeline implemented in PATRIC[38].	Table of relative abundances or bacterial genus.	A table of read counts per genus was first made for each sample. The total number of genus-aligned reads in that sample was then used to normalize the abundance of each genus.
Joining the tables of 16S abundances and shotgun abundances	Two relative abundance tables representing the genera present in each sample, one for 16S and one for shotgun.	A single table showing the relative abundances of each genus computed by the two methods.	The tables were joined using the logic of inner joining: only the genera present in both tables were included in the output table. Genera undetected in any of the samples (missing data) were set to 0.
Comparing shotgun and 16S at higher taxonomic levels	The genus-level table produced in the step above.	A table at each taxonomic level (family, order, class phylum)	The method aggregates the genus at each higher taxonomic level by adding relative abundances.

Table 1. Details for normalization and inclusion of taxa in the 16S vs. shotgun comparison.

Data Records

The shotgun sequenced samples were deposited in the NCBI/SRA as paired-end fastq files decontaminated of human reads[36,37]. The raw data can be found at *SRA* (the accession numbers for each of the 395 files are listed in 'AccessionShotgun') and *figshare*[44,45].

- samples.csv: We updated the data table tblASVsamples.csv[44] that we had previously published as part of our microbiota compilation[13]: We added a new column to the table, 'AccessionShotgun', which lists the SRA accession record for each of the 395 samples presented here. All other samples were left with an empty entry in column 'AccessionShotgun'. The table can be updated in the future as new shotgun sequences become available.
  - SampleID: stool sample identifier.
  - PatientID: deidentified identifier of patients.
  - Timepoint: deidentified day of sample collection.
  - Consistency: stool consistency.
  - Accession: the NCBI SRA accession number for the most recent submission (among all duplicate submissions) of the same 16S gene sequencing data corresponding to this sample.
  - BioProject: project-level SRA identifier for the chosen 'Accession'.
  - DayRelativeToNearestHCT: day of sample collection relative to the nearest day of bone marrow transplant.
  - AccessionShotgun: the NCBI SRA accession number for the shotgun sequencing of this sample.
  - Pool: The sequencing pool used for 16S samples.
  - Run: An identifier for each run used in 16S sequenced samples.
  - ShotgunBatchID: An identifier for each batch of shotgun-sequenced samples.

We compiled the additional tables for each sample as comma-separated value (csv) files[45] as following:

- ReadCounts.csv: list the 395 samples used this study for shotgun
  - SampleID: stool sample identifier.
  - Readcount: Number of reads for each sample after decontamination of human reads.

- Abundance: A Kraken 2 report provides information of the bacterial taxa in each sample.
  - Kindom, Phylum, Class, Order, Family, Genus: Each column contains name of taxonomic classification of each sample.
  - ColorOrder: Numeric data representing the order that each taxon was plotted in our manuscript.
  - HexColor: Color code in hex format for each taxon.
  - 395 columns using sample names as column names: Numeric data as the relative abundance of each sample. Every column should sum to 1.
    - NOTE: The (U)nclassified reads from kraken2 are not included in the calculation.
    - The sample names of those columns are modified to be compatible with the format requirement in matlab. To convert the names back to match the SampleID in the rest of the files, 's' at the beginning of each column name should be removed, and underscore ('_') needs to be converted to period (eg. 'sFMT_0001 A' will become 'FMT.0001A').

  - CARD.csv: A table provides information on genes known to confer antibiotic resistance in each sample.

    - Template: Hit of resistant genes in CARD.
    - Accession: NCBI accession number of the template.
    - Genome: Strain names where the template gene is found.
    - Species: The species of the strain.
    - resistGene: Gene name if available.
    - resistMechanism: Mechanism of resistance interpreted by CARD
    - Zoliflodacin-unknown (multiple columns): Antibiotics whether the gene is predicted to be resistant to, including unknown.

  - Score, Expected, Template_length, Template_Identity, Template_Coverage, Query_Identity, Query_Coverage, Depth, q_value, p_value: Parameters reported by CARD to show how well the match is.
  - shotgunReadcount: Number of reads for each sample after decontamination of human reads.
  - RelavantPercentInCARD: The number of reads matched to the template resistant gene/Total reads matched to all CARD genes in the sample.
  - PercentageInShotgun: The number of reads matched to the template resistant gene/Total reads in the sample.
  - Mutation: Information whether the antibiotic resistance is conferred by mutation
  - SampleID: Sample ID for each shotgun sequencing.

- VFDB.csv: A VFDB report provides information of the virulence factors in each sample.
  - Template: Hit of virulence genes in VFDB.
  - Function: Predicted function of the template.
  - Genome: Strain names where the template is found.
  - Score, Expected, Template_length, Template_Identity, Template_Coverage, Query_Identity, Query_Coverage, Depth, q_value, p_value: Parameters output by VFDB that report how well the match is.
  - shotgunReadcount: Number of reads for each sample after decontamination of human reads.
  - RelavantPercentInVF: The number of reads matched to the virulence gene/Total reads matched to virulence genes in the sample.
  - PercentageInShotgun: The number of reads matched to the virulence gene/Total reads in the sample.
  - SampleID: Sample ID for each shotgun sequencing.

- Taxa_Not_In_16S.csv: Taxa absent in 16S gene sequencing but present in shotgun metagenomic sequencing in our 395 samples.
  - TaxaName: Name of missing taxon.
  - Taxa: Classification level of taxa names (eg. Genus).
  - frequencyPresentInShotgun: Counts of non-zero abundance found in shotgun sequencing(non-zero)/Total number of samples.
  - medianRelAbdInShotgun: Median of relative frequency in shotgun metagenomic sequencing of that taxon.

- Taxa_Not_In_Shotgun.csv: Taxa absent in shotgun sequencing but present in 16S gene sequencing in our 395 samples.
  - TaxaName: Name of missing taxon.
  - Taxa: Classification level of taxon names (eg. Genus).
  - frequencyPresentIn16S: Counts of non-zero abundance found in 16S (non-zero)/total number of samples.
  - medianRelAbdIn16S: Median of relative frequency in 16S sequencing of that taxon.

Technical Validation

The nested subset of shotgun-sequenced samples explores various microbiome states experienced by patients. Out of the >10,000 samples with 16S rRNA gene sequencing[13], a total of 395 samples were sequenced using metagenomic shotgun technique for the purposes of different projects. Using a t-SNE map generated from Bray-Curtis dissimilarity matrix of 16S rRNA gene sequencing (Fig. 1a), we highlighted the samples with shotgun sequencing data available, which are distributed across the map (Fig. 1b). The nested subset captures a wide range of microbiome states, representing many states found in the original dataset of >10,000 samples. For example, both *Enterococcus*-dominated "dysbiotic" states (dark green portion of t-SNE projection in Fig. 1a) as well as the "healthier" Clostridia-enriched states (grey portion of t-SNE projection in Fig. 1a) are well-represented by the nested shotgun dataset.

The stool consistency from these patients varies widely. At the time of stool aliquoting, stool consistency was assessed by laboratory technicians using a scale of "formed, semi-formed, and liquid" to indicate the dry weight of stool[46]. The link between stool consistency and gut microbiota composition has been examined in the 16S amplicon sequencing pipeline[47]. Before diving into the diversity analysis, we first tested if the stool consistency would associate with the read count of shotgun sequencing. We observed that the median of the three types of stool (formed, semi-formed and liquid) are all above 10^7 reads per sample (Fig. 1c), except for two samples from the liquid group that showed lower reads count than the rest of the samples ($<10^5$).

Validating the taxonomic composition of the shotgun metagenomes. We first sought to compare the taxonomic classifications obtained by 16S rRNA gene sequencing[13] and shotgun sequencing analyzed by Kraken 2. A visual inspection of the bacterial compositions suggests that the two ways to analyze the taxonomic composition of the bacterial population agree well: When we compare the compositions from the patient with the highest number of collected samples we can see a reasonable match (Fig. 2a,b) between stacked bar plots of compositions color-coded according to a palette designed to highlight microbiome injury patterns[13].

A closer inspection shows however that the Shotgun sequencing missed some of the taxa seen in the shotgun data (eg. the orange bar representing *Erysipelotrichia* in day 56, and the blue bar representing *Bacilli* in day 82). We then compared the relative abundance of different taxa as assessed by 16S and shotgun sequencing and identified the taxa with median relative abundance higher than 10% and significantly different between 16S and shotgun sequencing, among which the Firmicutes (phylum) has the overall highest abundances (Fig. 2c), which could be

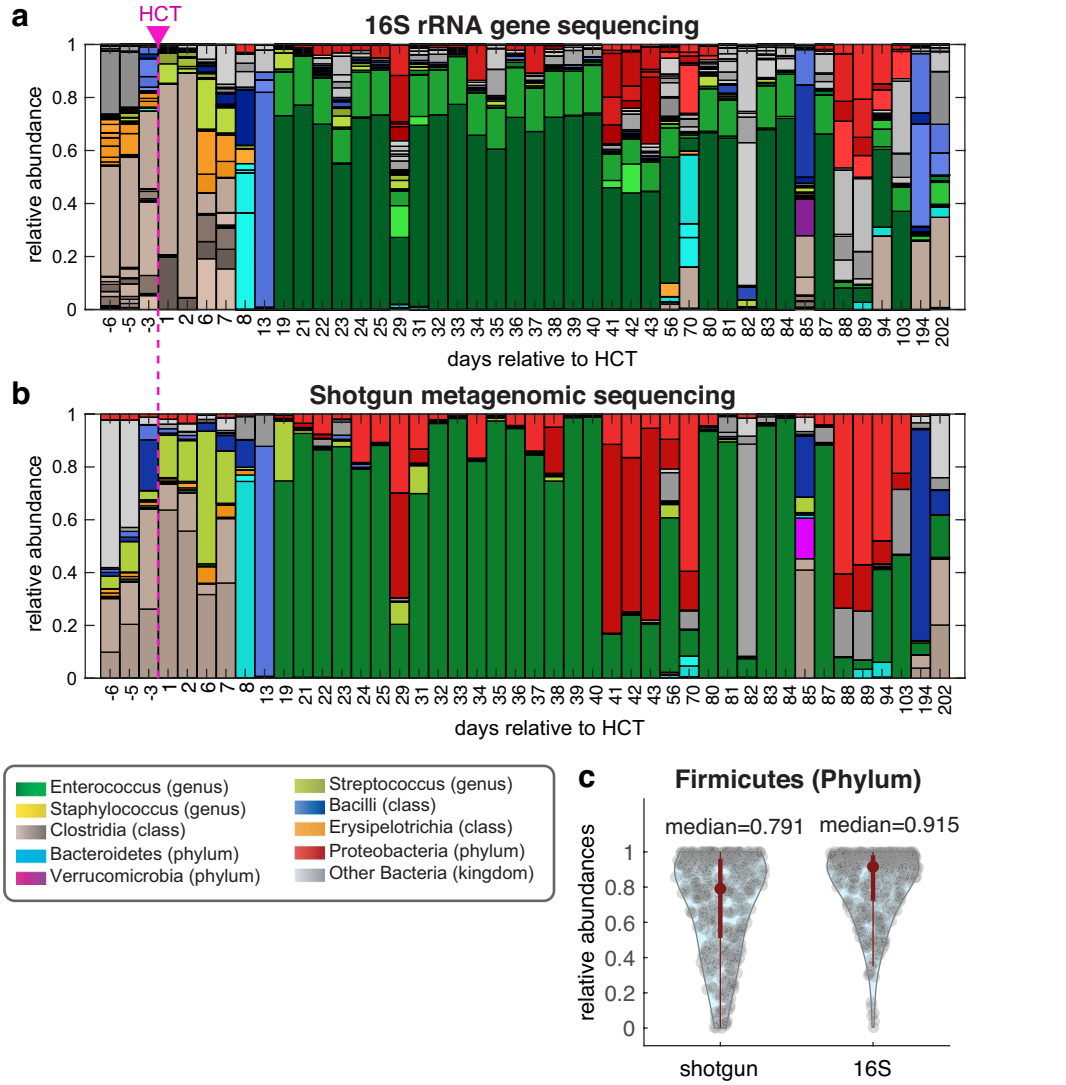

Fig. 2 Taxonomic composition of the microbiome in patient stool samples agrees in general between shotgun sequencing and 16S rRNA amplicon sequencing, with some notable differences. (**a**,**b**) The taxonomic composition is determined by 16S rRNA sequencing (A) and shotgun metagenomics (B) for the samples from a single patient (PatientID 1252). The samples are ordered in time and the dashed line separates the samples collected before and after allo-HCT. (**c**) The median composition (red dot) in Firmicutes can be notably different when determined using the two approaches (ranksum test, $p < 0.05$).

explained by its higher copy number of rRNA in the genome[48]. The other taxonomic groups are Bacilli (class), Clostridia (class), Clostridiales (order) and Lactobacillales (order).

There were some taxa only found in either approach, and the shotgun sequencing found much more taxa than the 16S gene sequencing (1870 missing in 16S but present in shotgun; 182 missing in shotgun but present in 16S; the.csv files can be found in Figshare). There are a few reasons that could possibly explain the disagreement between 16S and metagenomic shotgun sequencing: First is the ambiguous naming where a taxon was renamed later (e.g. Enterobacteriales was renamed to Enterobacterales), sharing the same sequencing of tested 16S region (Escherichia and Shigella are the same in 16S gene sequencing), and poorly studied taxa (e.g. 'CAG-352' in 16S sequencing). The second is the detection variation. The taxa missing entirely either in 16S or in shotgun overall have very low median abundance even detected in the other pipeline – only 'Incertae Sedis' and 'CAG-352' found in 16S are between 1 to 4 percent, while all other missing taxa show less than 1% median relative abundance in the other pipeline where they are found. The third reason could be differences between the databases used for taxonomy detection. To systematically compare 16S and shotgun sequencing, we calculated correlation of relative abundance between the two pipelines in taxonomic classification. The correlation method was the Pearson correlation between the two estimates of each taxon's relative abundance in the same sample: the value obtained from 16S analysis and the value obtained from shotgun analysis. The agreement between 16S and metagenomic shotgun were generally high and decreased for lower taxonomic ranks (Fig. 3), indicating that

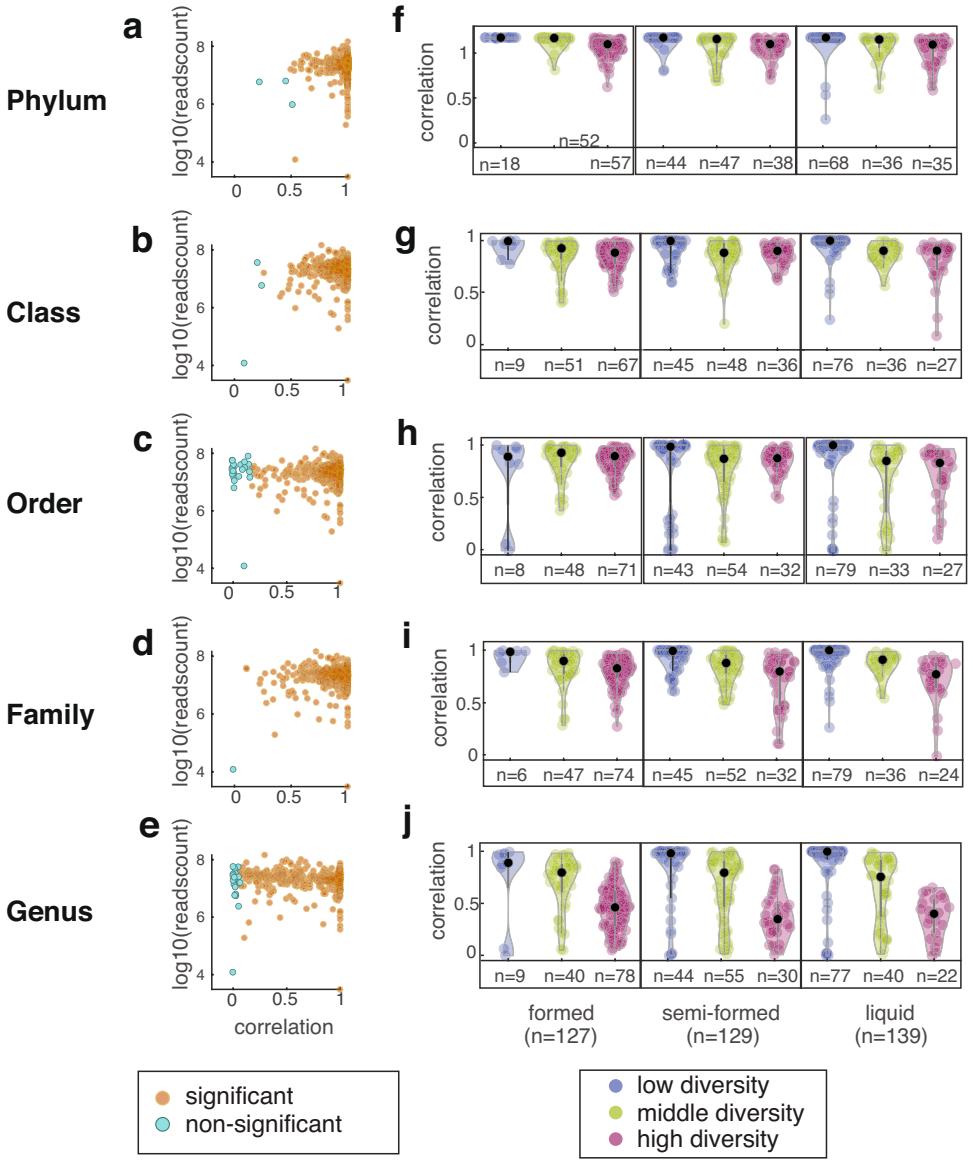

Fig. 3 Correlation between the taxonomic classifications obtained by shotgun sequencing and 16S rRNA amplicon sequencing. The correlation between the two approaches is different at each taxonomic level, but seems unaffected by the read depth of each sample (**a**–**e**). Formed stool mainly contains samples with higher diversity, and high diversity samples usually display lower correlation between the two sequencing pipelines (**f**–**j**). Each point is a taxon from one of 395 samples. Black dots in (**f**–**j**) indicate the median of each category. The numbers on the x-axes display the number of samples in different diversity groups.

the two approaches have different sensitivities for taxonomy. The few samples with low correlation tended to be those sequenced at a lower read depth, but some samples sequenced deeply could also have low correlation (Fig. 3a–e), indicating that other factors than read counts may affect the taxonomic mapping of metagenomes.

One possible explanation is that discrepancies between the database of bioinformatic pipelines may become especially visible for highly diverse samples, leading to low correlations. We therefore calculated the alpha-diversity as determined by the Shannon index for each taxonomic classification. Then we divided the values into three groups: high diversity (top 33%), middle diversity and low diversity (bottom 33%). Because stool consistency is a marker of species richness in microbiome[47], we stratified our samples by stool consistency and examined if the diversity clusters are discrete among different consistencies (Fig. 3f–j). The high diversity group does have overall lower correlation, which becomes more and more obvious from phylum to genus. We also noted 5 samples with both low diversity (in bottom 33% percentile) and low correlation (less than 0.01) between the taxonomies quantified by shotgun and 16S. Four samples with the lowest correlation (<0.001) in genus composition are also in the bottom 33% percentile of diversity, which were caused by a failure by the 16S pipeline to detect a bacterium of the genus *Klebsiella*. This example illustrates a possible source of error in the 16S pipeline that may be improved using shotgun metagenomics.

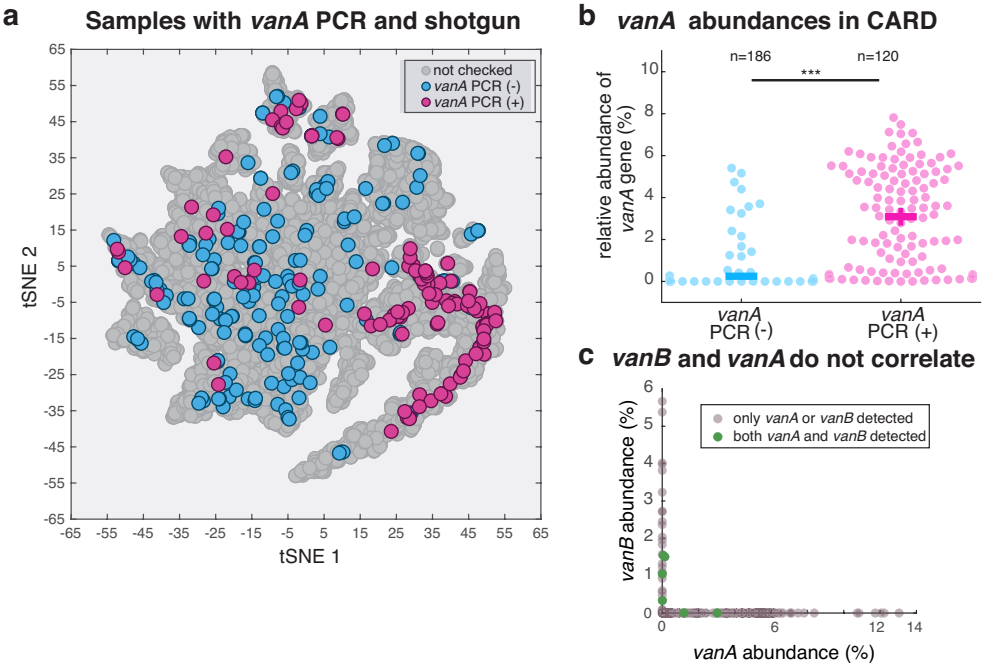

Fig. 4 The shotgun sequencing detects the presence of antibiotic resistance genes, using the PATRIC service with the CARD database. (**a**) Localization of the *vanA*(+/−) samples in 16S clustering map shows a high concentration of *vanA*(+) samples in the region of domination by *Enterococcus* (green in Fig. 1a). (**b**) PCR(+) samples have higher relative abundance of the *vanA* gene detected by shotgun sequencing. (**c**) The *vanA* and *vanB* genes are practically mutually exclusive in patients' stool samples. The samples with two genes simultaneously detected represent a very small fraction of the total samples. The abundances of the two genes are not correlated.

Validating the detection of antibiotic resistance genes using a PCR to detect the *vanA* gene. PATRIC provides web services to quantify virulence factors and antibiotic resistance genes in the microbiome samples. To test how well this analysis detected antibiotic resistant genes, we used PCR to detect the presence of an important gene for vancomycin resistance *vanA*[13] (Fig. 4a). Vancomycin is a glycopeptide antibiotic that is given to many of the allo-HCT patients in this cohort as prophylaxis to prevent infections by *Streptococcus*[49]. The samples chosen for the PCR test contain ≥2% enterococcal sequences, since the presence of *vanA* is usually a sign of *Enterococcus* domination[50,51]. We compared the relative abundance of the vanA gene, as quantified by the PATRIC analysis of CARD genes, in *vanA* PCR(-) versus *vanA* PCR(+) samples and we saw a significant agreement (Fig. 4b). In the PCR(+) group only 2 samples (out of 120) have zero abundance of *vanA* in metagenomes while 143 out of 186 are zero in PCR(-) group, suggesting that shotgun sequencing may be more sensitive than the PCR.

There are other genes besides *vanA* important for resistance to vancomycin. We examined whether the abundance of another vancomycin resistance gene, *vanB*, correlated with *vanA*. We saw that although those two genes are negatively correlated in our gut microbiome samples ($r = -0.28$, $p < 0.05$), plotting the gene abundance (Fig. 4c) reveals that only six samples carry both *vanA* and *vanB* (Fig. 4c, green dots). In all the other cases, only *vanA* or *vanB* was found, suggesting that bacteria harboring these genes may be excluding invasion by competitors harboring the other gene.

Validating the assembly of genomes from shotgun sequences. New bioinformatic pipelines have enabled us to assemble the genomes of bacteria from shotgun sequences[40,52,53]. To illustrate the utility of our data for this type of analysis, we ran a published metagenomic analysis pipeline to find MAGs (metagenome-assembled genomes) from the samples of PatientID 1044, where we know from genotyped isolates obtained in a previous study that the patient carried *Enterococcus faecium* in the gut[34]. We found 7 high-quality MAGs classified as *E. faecium*, each from a different stool sample and has a completeness higher than 95%. We then compared these MAGs with the genomes of our 26 *E. faecium* isolates[34] in a phylogenetic tree (Fig. 5). Our previous study[34] had shown that the patient contained at least two distinct strains of *E. faecium*. The MAGs confirmed the observation: Three MAGS, MAG_1044M_maxbin0003, MAG_1044P_maxbin002 and MAG_1044L_4, located in the tree branch of the strains from the same three samples that collected in the later days relevant to HCT, whereas four MAGs, MAG_1044J_16, MAG_1044G_10, MAG_1044H_27 and MAG_1044I_maxbin001, located in the tree branch of the other strains that were isolated from the same four samples from the earlier days after HCT. The comparative analysis indicates that the dominant *E. faecium* strain has shifted between day 5 and day 7 after HCT.

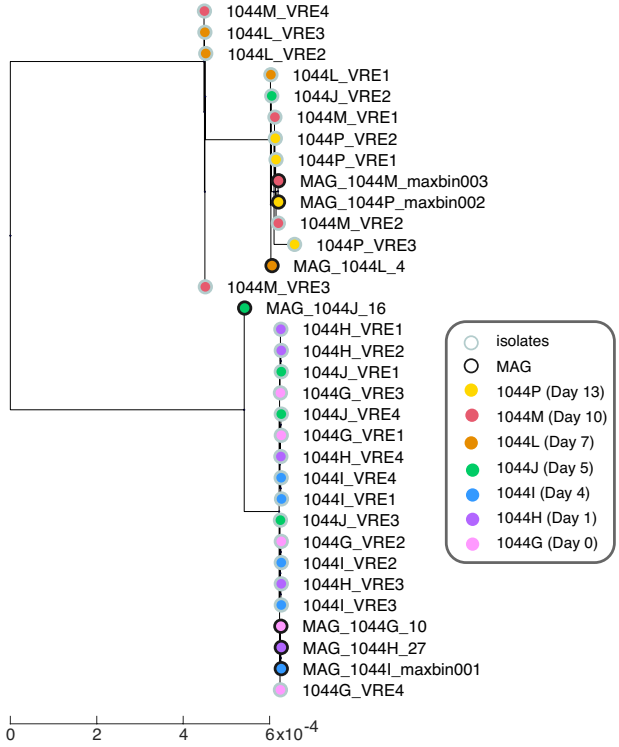

Fig. 5 Shotgun sequencing data provide metagenomically-assembled genomes (MAGs) that compare well with the genomes of isolates from the same patient stool samples. MAGs from *E. faecium* obtained from different samples collected from patient 1044 reveal an intraspecies diversity. The phylogenetic tree contains the 7 MAGs and 26 *E. faecium* genomes obtained from isolates and analyzed in a previous study[34]. The number of days after each sample is the day relative to the HCT of this patient.

Code availability

The analysis code (Matlab 2020a) used for the examples provided below is available in the GitHub repository https://github.com/Jinyuan1998/scientific_data_metagenome_shotgun. The script used for each figure is in a separate directory:

- Figure 1: Figuer1/scFigure 1.m.
- Figure 2: Figuer2/scFigure 2.m.
- Figure 3: Figuer3/scFigure 3.m.
- Figure 4: Figuer4/scFigure 4.m.
- Figure 5: Figuer5/scFigure 5.m.

Received: 19 October 2021; Accepted: 21 February 2022;
Published online: 18 May 2022

References

1. David, L. A. *et al.* Diet rapidly and reproducibly alters the human gut microbiome. *Nature* **505**, 559–563 (2014).
2. Papanicolas, L. E. *et al.* Conventional myelosuppressive chemotherapy for non-haematological malignancy disrupts the intestinal microbiome. *BMC Cancer* **21**, 591 (2021).
3. Dethlefsen, L. & Relman, D. A. Incomplete recovery and individualized responses of the human distal gut microbiota to repeated antibiotic perturbation. *Proc. Natl. Acad. Sci. USA* **108**(Suppl 1), 4554–4561 (2011).
4. Integrative HMP (iHMP) Research Network Consortium. The integrative human microbiome project. *Nature* **569**, 641–648 (2019).
5. Wlodarska, M., Kostic, A. D. & Xavier, R. J. An integrative view of microbiome-host interactions in inflammatory bowel diseases. *Cell Host Microbe* **17**, 577–591 (2015).
6. Kinross, J. M., Darzi, A. W. & Nicholson, J. K. Gut microbiome-host interactions in health and disease. *Genome Med.* **3**, 14 (2011).
7. Stoma, I. *et al.* Compositional flux within the intestinal microbiota and risk for bloodstream infection with gram-negative bacteria. *Clin. Infect. Dis.* https://doi.org/10.1093/cid/ciaa068 (2020).
8. Morjaria, S. *et al.* Antibiotic-Induced Shifts in Fecal Microbiota Density and Composition during Hematopoietic Stem Cell Transplantation. *Infect. Immun.* **87**, (2019).
9. Golob, J. L. *et al.* Stool microbiota at neutrophil recovery is predictive for severe acute graft vs host disease after hematopoietic cell transplantation. *Clin. Infect. Dis.* **65**, 1984–1991 (2017).
10. Ferrer, M., Méndez-García, C., Rojo, D., Barbas, C. & Moya, A. Antibiotic use and microbiome function. *Biochem. Pharmacol.* **134**, 114–126 (2017).
11. Willing, B. P., Russell, S. L. & Finlay, B. B. Shifting the balance: antibiotic effects on host-microbiota mutualism. *Nat. Rev. Microbiol.* **9**, 233–243 (2011).
12. Schluter, J. *et al.* The gut microbiota is associated with immune cell dynamics in humans. *Nature* **588**, 303–307 (2020).

13. Liao, C. *et al*. Compilation of longitudinal microbiota data and hospitalome from hematopoietic cell transplantation patients. *Sci. Data* **8**, 71 (2021).
14. Stein-Thoeringer, C. K. *et al*. Lactose drives Enterococcus expansion to promote graft-versus-host disease. *Science* **366**, 1143–1149 (2019).
15. Taur, Y. *et al*. Reconstitution of the gut microbiota of antibiotic-treated patients by autologous fecal microbiota transplant. *Sci. Transl. Med*. **10** (2018).
16. Markey, K. A. *et al*. The microbe-derived short-chain fatty acids butyrate and propionate are associated with protection from chronic GVHD. *Blood* **136**, 130–136 (2020).
17. Peled, J. U. *et al*. Microbiota as Predictor of Mortality in Allogeneic Hematopoietic-Cell Transplantation. *N. Engl. J. Med.* **382**, 822–834 (2020).
18. Zhai, B. *et al*. High-resolution mycobiota analysis reveals dynamic intestinal translocation preceding invasive candidiasis. *Nat. Med.* **26**, 59–64 (2020).
19. Haak, B. W. *et al*. Impact of gut colonization with butyrate-producing microbiota on respiratory viral infection following allo-HCT. *Blood* **131**, 2978–2986 (2018).
20. Peled, J. U. *et al*. Intestinal Microbiota and Relapse After Hematopoietic-Cell Transplantation. *J. Clin. Oncol.* **35**, 1650–1659 (2017).
21. Shono, Y. *et al*. Increased GVHD-related mortality with broad-spectrum antibiotic use after allogeneic hematopoietic stem cell transplantation in human patients and mice. *Sci. Transl. Med.* **8**, 339ra71 (2016).
22. Jenq, R. R. *et al*. Intestinal Blautia Is Associated with Reduced Death from Graft-versus-Host Disease. *Biol. Blood Marrow Transplant.* **21**, 1373–1383 (2015).
23. Hillmann, B. *et al*. Evaluating the information content of shallow shotgun metagenomics. *mSystems* **3**, (2018).
24. Qin, J. *et al*. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464**, 59–65 (2010).
25. Knight, R. *et al*. Best practices for analysing microbiomes. *Nat. Rev. Microbiol.* **16**, 410–422 (2018).
26. Wood, D. E., Lu, J. & Langmead, B. Improved metagenomic analysis with Kraken 2. *Genome Biol.* **20**, 257 (2019).
27. Alcock, B. P. *et al*. CARD 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database. *Nucleic Acids Res.* **48**, D517–D525 (2020).
28. Liu, B., Zheng, D., Jin, Q., Chen, L. & Yang, J. VFDB 2019: a comparative pathogenomic platform with an interactive web interface. *Nucleic Acids Res.* **47**, D687–D692 (2019).
29. Kang, D. D. *et al*. MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ* **7**, e7359 (2019).
30. Alneberg, J. *et al*. Binning metagenomic contigs by coverage and composition. *Nat. Methods* **11**, 1144–1146 (2014).
31. Abubucker, S. *et al*. Metabolic reconstruction for metagenomic data and its application to the human microbiome. *PLoS Comput. Biol.* **8**, e1002358 (2012).
32. Belcour, A. *et al*. Metage2Metabo, microbiota-scale metabolic complementarity for the identification of key species. *Elife* **9** (2020).
33. Pascal Andreu, V., Roel-Touris, J., Dodd, D., Fischbach, M. A. & Medema, M. H. The gutSMASH web server: automated identification of primary metabolic gene clusters from the gut microbiota. *Nucleic Acids Res.* **49**, W263–W270 (2021).
34. Dubin, K. A. *et al*. Diversification and Evolution of Vancomycin-Resistant Enterococcus faecium during Intestinal Domination. *Infect. Immun*. **87** (2019).
35. Nowinski, B. *et al*. Microbial metagenomes and metatranscriptomes during a coastal phytoplankton bloom. *Sci. Data* **6**, 129 (2019).
36. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/bioproject:PRJNA545312 (2019).
37. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/bioproject:PRJNA607574 (2020).
38. Davis, J. J. *et al*. The PATRIC Bioinformatics Resource Center: expanding data and analysis capabilities. *Nucleic Acids Res.* **48**, D606–D612 (2020).
39. Clausen, P. T. L. C., Aarestrup, F. M. & Lund, O. Rapid and precise alignment of raw reads against redundant databases with KMA. *BMC Bioinformatics* **19**, 307 (2018).
40. Siranosian, B. A. *et al*. Rare transmission of commensal and pathogenic bacteria in the gut microbiome of hospitalized adults. *BioRxiv* https://doi.org/10.1101/2021.03.12.435204 (2021).
41. Nurk, S., Meleshko, D., Korobeynikov, A. & Pevzner, P. A. metaSPAdes: a new versatile metagenomic assembler. *Genome Res.* **27**, 824–834 (2017).
42. 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).
43. Sieber, C. M. K. *et al*. Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. *Nat. Microbiol.* **3**, 836–843 (2018).
44. Liao, C. The day of collection of 12,659 stool samples for 1,879 patients and the stool consistency. *Figshare* https://doi.org/10.6084/m9.figshare.12016983 (2020).
45. Yan, J. & Xavier, J. A compilation of fecal microbiome shotgun metagenomics from hematopoietic cell transplantation patients. *Figshare* https://doi.org/10.6084/m9.figshare.c.5623885.v2 (2021).
46. Vork, L. *et al*. Does Day-to-Day Variability in Stool Consistency Link to the Fecal Microbiota Composition? *Front. Cell Infect. Microbiol*. **11** (2021).
47. Vandeputte, D. *et al*. Stool consistency is strongly associated with gut microbiota richness and composition, enterotypes and bacterial growth rates. *Gut* **65**, 57–62 (2016).
48. Větrovský, T. & Baldrian, P. The variability of the 16S rRNA gene in bacterial genomes and its consequences for bacterial community analyses. *PLoS One* **8**, e57923 (2013).
49. Seo, S. K. *et al*. Impact of peri-transplant vancomycin and fluoroquinolone administration on rates of bacteremia in allogeneic hematopoietic stem cell transplant (HSCT) recipients: a 12-year single institution study. *J. Infect.* **69**, 341–351 (2014).
50. Taur, Y. *et al*. Intestinal domination and the risk of bacteremia in patients undergoing allogeneic hematopoietic stem cell transplantation. *Clin. Infect. Dis.* **55**, 905–914 (2012).
51. Ubeda, C. *et al*. Vancomycin-resistant Enterococcus domination of intestinal microbiota is enabled by antibiotic treatment in mice and precedes bloodstream invasion in humans. *J. Clin. Invest.* **120**, 4332–4341 (2010).
52. Albanese, D. & Donati, C. Strain profiling and epidemiology of bacterial species from metagenomic sequencing. *Nat. Commun.* **8**, 2260 (2017).
53. Olm, M. R. *et al*. inStrain profiles population microdiversity from metagenomic data and sensitively detects shared microbial strains. *Nat. Biotechnol.* **39**, 727–736 (2021).

Acknowledgements

This work was supported by the National Institutes of Health (NIH) grants U01 AI124275 to EP and JBX and R01 AI137269 to JBX and YT. This work could not have been possible without our longtime collaborator Eric Littman. Eric passed away in June 2021 when we were starting to prepare this manuscript. We acknowledge Eric's invaluable contribution to this dataset and his analyses of shotgun sequencing data that inspired some of the examples presented here.

Author contributions

J.Y., C.L. and J.B.X. wrote the manuscript. J.Y., C.L. and J.B.X. designed the analyses. Y.T. contributed to the clinical data preparation, B.P.T. and E.R.L provided the 16S data processing pipelines; together with A.D., N.W. and J.U.P. they did some of the preliminary analysis that led to this work. B.AS. and A.S.B. provided the pipelines for strain-level analysis and interpretation of the results. E.F., L.A.A. and R.J.W. processed patients' stool samples. M.A.P., M.R.M.B., E.G.P., J.S. and J.B.X. led the project. All authors contributed to the writing and interpretation of the results.

Competing interests

M.R.M.v.d.B. received financial support from Seres Therapeutics. J.U.P. reports research funding, intellectual property fees, and travel reimbursement from Seres Therapeutics, and consulting fees from DaVolterra, CSL Behring, and from MaaT Pharma. He has filed intellectual property applications related to the microbiome (reference numbers #62/843,849, #62/977,908, and #15/756,845). M.-A.P/ has received honoraria from AbbVie, Bellicum, Bristol-Myers Squibb, Incyte, Merck, Novartis, Nektar Therapeutics, and Takeda; has received research support for clinical trials from Incyte, Kite (Gilead) and Miltenyi Biotec; and serves on data and safety monitoring boards for Servier and Medigene and scientific advisory boards for MolMed and NexImmune.

Additional information

Correspondence and requests for materials should be addressed to J.Y. or J.B.X.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.