nature biotechnology

ARTICLES

https://doi.org/10.1038/s41587-021-01102-3

Fast and accurate metagenotyping of the human gut microbiome with GT-Pro

Zhou Jason Shi[1,2], Boris Dimitrov[3], Chunyu Zhao[1], Stephen Nayfach[4,5] ✉ and Katherine S. Pollard[1,2,6] ✉

Single nucleotide polymorphisms (SNPs) in metagenomics are used to quantify population structure, track strains and identify genetic determinants of microbial phenotypes. However, existing alignment-based approaches for metagenomic SNP detection require high-performance computing and enough read coverage to distinguish SNPs from sequencing errors. To address these issues, we developed the GenoTyper for Prokaryotes (GT-Pro), a suite of methods to catalog SNPs from genomes and use unique *k*-mers to rapidly genotype these SNPs from metagenomes. Compared to methods that use read alignment, GT-Pro is more accurate and two orders of magnitude faster. Using high-quality genomes, we constructed a catalog of 104 million SNPs in 909 human gut species and used unique *k*-mers targeting this catalog to characterize the global population structure of gut microbes from 7,459 samples. GT-Pro enables fast and memory-efficient metagenotyping of millions of SNPs on a personal computer.

Microbial species harbor extensive genetic variation, including SNPs, structural variants and mobile genetic elements. SNPs in particular are useful for population genetic analyses[1], such as tracking transmission of strains between environments or locations, reconstructing intraspecies phylogenetic relationships, resolving mixtures of genotypes within a host and depicting population diversity or structure along environmental gradients. Additionally, SNPs can result in or be linked to changes in protein function and microbial phenotypes[2–4]. Being able to broadly and accurately quantify intraspecies genomic variation in the human microbiome is a prerequisite to the potential application of microbiome genomics to precision medicine.

The gold standard approach for identifying SNPs in microbiomes is to sequence individual isolate genomes and identify mismatches in whole-genome alignments[5]. Another approach is to align short metagenomic reads to reference genomes, which circumvents the need for strain isolation. This 'metagenotyping' strategy was implemented by Schloissnig et al.[6] who discovered 10.3 million SNPs for 101 human gut species and has since been featured in several tools, including Constrains[7], MIDAS[8], metaSNV[9], DESMAN[10] and StrainPhlAn[11]. While algorithms for read alignment have improved, the approach is still computationally costly, requires enough coverage to distinguish between SNPs and sequencing errors, and depends on large databases of microbial genomes. Exact matching algorithms such as Kraken[12], CLARK[13] and bfMEM[14] can process reads orders of magnitude faster than alignment but have not been used to perform SNP calling and can suffer from false positives where short sequences (*k*-mers) occur in multiple taxa[15]. This problem has been addressed in microbial forensics by Insignia[16], which compares bacterial and viral genomes to identify DNA 'signatures' that are unique to a target of interest, and KrakenUniq[15] that uses unique *k*-mers for taxonomic classification.

Inspired by these techniques, we sought to deploy a *k*-mer exact matching algorithm to rapidly and accurately genotype SNPs using shotgun metagenomics data. We were motivated by the LAVA method[17], which uses *k*-mer exact matching to one allele or the other for genotyping known biallelic SNPs in human whole-genome sequencing data. We hypothesized that the LAVA genotyping strategy could be extended to metagenomics by optimizing ideas from taxonomic classifiers such as Insignia and KrakenUniq. We had to solve three main problems. First, we compiled high-resolution genetic maps to identify the location of biallelic SNPs in conserved genomic regions for hundreds of microbiome species. Second, we created an in silico metagenotyping array that contained *k*-mers capable of uniquely probing each allele of every SNP. Third, we developed scalable algorithms and data structures to rapidly and efficiently search the billions of *k*-mers in this array against millions of sequencing reads from a typical metagenome. Our goal was to develop a metagenotyping software tool that is at least as accurate as read alignment methods, while being computationally efficient enough to run on a personal computer.

Results

A framework for in silico genotyping of microbiome species. We introduce the GenoTyper for PROkaryotes (GT-Pro), which is an open-source software suite, to perform fast and accurate metagenotyping (Fig. 1). The key components of GT-Pro are (1) a compact data structure encoding SNP-covering *k*-mers (sck-mers) that captures most common variation found in genomes from an environment of interest, (2) a procedure for selecting highly species-specific sck-mers to reduce false positive metagenotypes and (3) a metagenotyping algorithm that combines and optimizes hashing, filtering and data compression for exact matching of species-specific sck-mers to *k*-mers in shotgun metagenomes. Building a version of GT-Pro for a given environment involves using reference genomes or metagenome-assembled genomes (MAGs) to discover common SNPs in conserved regions for each species and selecting species-specific sck-mers to include on the metagenotyping array. We focus on common SNPs because this results in a data structure small enough to fit in computer memory while still capturing most genetic variation for most species. Building the array for a new environment may require high-performance computing,

[1]Data Science, Chan Zuckerberg Biohub, San Francisco, CA, USA. [2]Data Science and Biotechnology, Gladstone Institutes, San Francisco, CA, USA. [3]Chan Zuckerberg Initiative, Redwood city, CA, USA. [4]Department of Energy, Joint Genome Institute, Walnut Creek, CA, USA. [5]Environmental Genomics and Systems Biology Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA. [6]Epidemiology and Biostatistics, University of California, San Francisco, CA, USA. ✉e-mail: snayfach@lbl.gov; katherine.pollard@gladstone.ucsf.edu

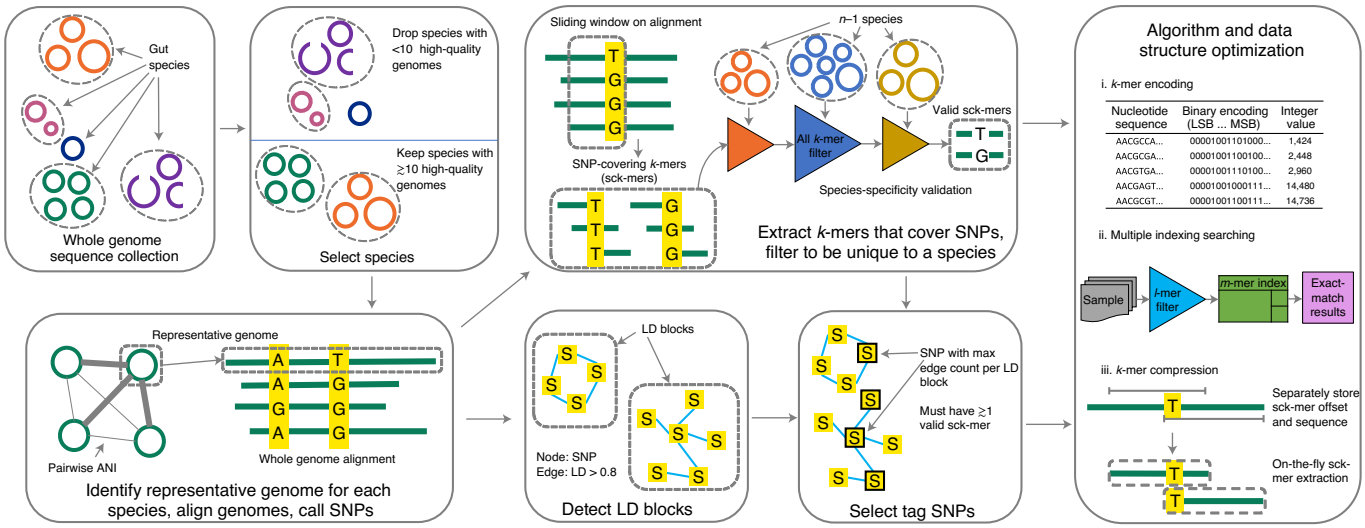

Fig. 1 | In sillico metagenotyping framework. GT-Pro starts with a whole-genome sequence collection and identifies species with sufficient high-quality genomes to call SNPs. For each species, a representative genome is chosen based on pairwise average nucleotide identity (ANI) plus assembly quality metrics. SNPs are called per species based on whole-genome alignment of conspecific genomes to the representative genome. Common (site prevalence ≥90% and minor allele frequency >1%) biallelic SNPs are selected for genotyping. Up to four times *k* candidate *k*-mers are extracted per SNP site, covering both the reference and alternative allele on forward and reverse complementary strands (sck-mers, $k=31$ in this study). These candidate sck-mers are iteratively filtered through species-specificity filters of all unique *k*-mers present in the genomes of every other species, not including species with insufficient high-quality genomes for genotyping. Only SNPs with sck-mers for both the reference and alternative allele are retained. Next, SNPs are clustered based on co-occurence patterns across genomes into linkage disequilibrium (LD) blocks. LD blocks are detected with an R^2 threshold (0.81), and a tag SNP with species-specific sck-mers and the highest LD to other SNPs in the block is selected. Optimized algorithms and compressed representations of sck-mer data enable rapid metagenotyping. Further details are shown in the Methods and Fig. 3. max, maximum.

but running GT-Pro on microbiome samples does not. As a proof of principle, we applied GT-Pro to the human gut microbiome. We reasoned that, given the large number of sequenced gut genomes, we would be able to build high-resolution genetic maps for many species that would allow us to accurately and comprehensively construct a metagenotyping array.

Using 112,904 high-quality genomes (≥90% completeness and ≤5% contamination[18]), we identified SNPs for 909 human gut species (minimum, ten genomes; median, 35 genomes) (Supplementary Figs. 1 and 2 and Supplementary Table 1). These include both MAGs[19–21] (94.1%) and cultivated isolates (5.9%), and were derived from geographically and phenotypically diverse human participants. We both identified representative genomes and performed whole-genome alignments for each species using MUMmer4 (ref. [22]), revealing 104,171,172 common, core-genome SNPs (minor allele frequency ≥1%, site prevalence ≥90%). Most SNPs were biallelic (Fig. 2a, Supplementary Fig. 3a and Supplementary Fig. 4) and an extremely low fraction disrupted a stop codon or introduced a premature one, which is one indicator of false positives (Fig. 2a). For context, this catalog is tenfold larger than the one established by Schloissnig et al.[6] and 1.22-fold larger than the catalog of all human SNPs[23] (Supplementary Fig. 1). Consistent with previous reports[6], SNP density, nucleotide diversity and the rate of nonsynonymous versus synonymous mutations (pN/pS) varied across species and phyla (Fig. 2b and Supplementary Figs. 5–9), which may reflect differences in selective pressures, population sizes or transmission modes.

Species-specific *k*-mers enable accurate and efficient identification of SNPs. Having constructed a large SNP catalog of the gut microbiome, we next constructed a metagenotyping array that contained sck-mers that could uniquely identify each SNP from a shotgun metagenome. Similar to both Kraken ($k=31$) and LAVA ($k=32$), we chose a length of $k=31$ to ensure high specificity across the gut microbiome while limiting compute and memory requirements. Of the roughly 12.9 billion candidate 31-mers that overlapped a SNP (124 per SNP; 31 per allele type and sequence orientation), we identified 5.7 billion that were unique. These species-specific sck-mers overlapped 97% of the 909 species and 51% of the 104 million SNPs (mean 108 sck-mers per SNP, Supplementary Figs. 1 and 10). The species that cannot be genotyped with this strategy due to insufficient sck-mers tend to have a very close relative. These are most common within *Actinobacteria* (Fig. 2c and Supplementary Fig. 3b). Our sck-mers capture 83% of the within-species variation compared to whole-genome average nucleotide identity, and they provide a much higher level of resolution compared to using genetic variation in phylogenetic marker genes (16S or universal, single-copy proteins; Supplementary Fig. 11).

Compact storage of sck-mers in computer memory. To efficiently fit the GT-Pro database in memory, we implemented a data structure that separately stores a 60-basepair (bp) sequence centered on each SNP and an index of positions at which sck-mers occur in the sequence (Methods). This requires only 13 GB of RAM and permits GT-Pro to run on most modern personal computers (Supplementary Figs. 12 and 13). Storing one sequence window for each SNP is efficient, because 98.2% of SNPs are separated from the closest other SNP by at least 30-bp so most windows are nonoverlapping.

To further reduce the database size, we used single-linkage clustering to group the 104 million SNPs into only 6.8 million physically linked blocks ($R^2>0.81$) that covaried across reference genomes (Supplementary Fig. 14) and selected a single tag SNP per block. This represents a >15-fold reduction in database size. A similar strategy is commonly used when designing genotyping chips. Although recombination mechanisms in bacteria are distinct from those in animals, most gut microbiome species are not clonal and their genomes show varying degrees of linkage disequilibrium (LD)[24–29]. Our choice of R^2 is motivated by thresholds used for high-confidence SNP imputation in other species and the fact

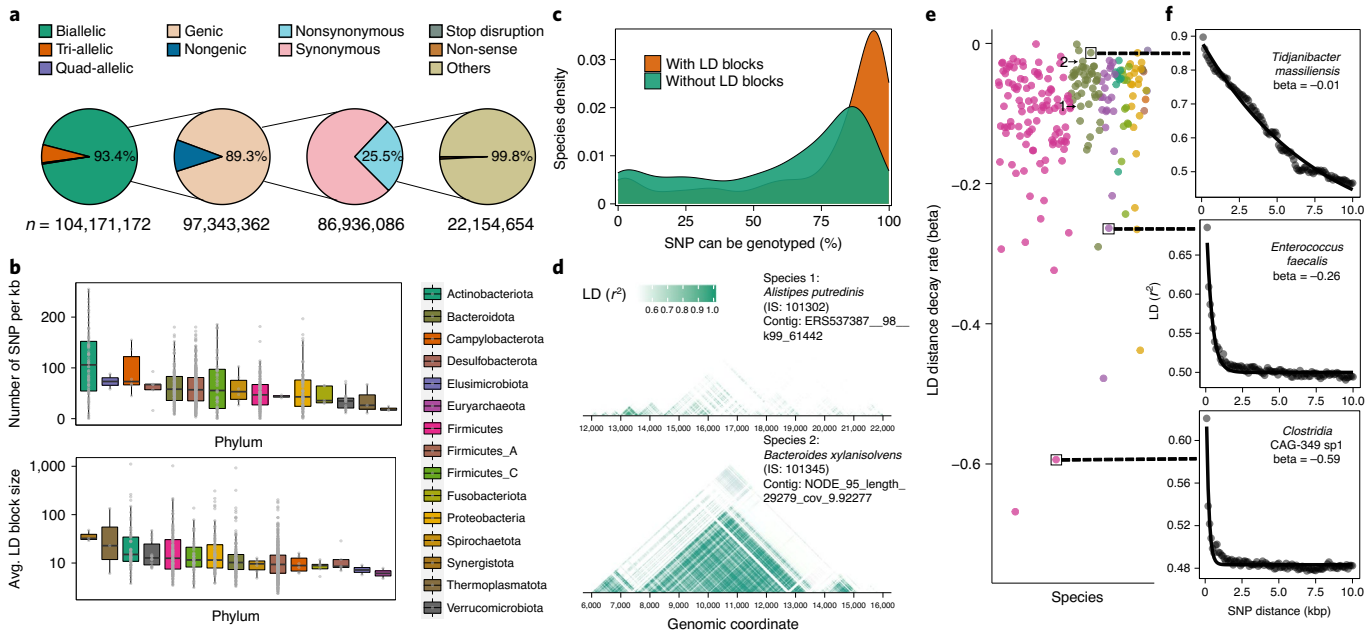

Fig. 2 | Genetic landscape of 909 human gut species. a, Summary of common SNP characteristics across all species (from left to right): at most SNPs, only two alleles are observed, biallelic SNPs are mostly within protein-coding genes, these are largely synonymous and the nonsynonymous ones rarely disrupt or introduce a stop codon. **b**, Phyla differ in their median SNP density (upper) and average LD block size (lower) with sizeable variation in density across species within each phylum. A standard boxplot is used here and elsewhere. Avg., average. **c**, Distribution across species of the percentage of common SNPs that can be genotyped by GT-Pro either directly ('without LD blocks') or are in an LD block with a tag SNP that can be genotyped ('with LD blocks'). For a typical species, roughly 75% of SNPs can be genotyped directly and roughly 95% are physically linked to a tag SNP that can be genotyped. **d**, Visualization of two distinct haplotype landscapes from (upper) *Alistipes putredinis* (species ID 101302) and (lower) *Bacteroides xylanisolvens* (species ID 101345), both with fairly high LD compared to other species. Horizontal axis is genomic coordinate. Color indicates magnitude of LD between pairs of SNPs. The examples have the same genomic span (10,000 bp). **e**, Rate of LD distance decay across gut bacterial species with ≥100 genomes ($n = 228$). Same phylum color scheme as **b**. Black arrow points to the species in **d**: 1 is *Alistipes putredinis* and 2 is *Bacteroides xylanisolvens*. **f**, Examples of LD distance decay for individual species. From top to bottom are three species (species IDs 102371, 101694 and 102831) with increasing LD distance decay, suggesting higher recombination rates. Curves represent the fitted exponential decay model.

that discovery of LD blocks stabilizes in this range for gut species (Supplementary Fig. 15). On average LD blocks spanned roughly 4.3 kilobasepairs (kbp) and 23.5 SNPs, although the number and size of LD blocks varied considerably across bacterial species (Fig. 2b and Supplementary Figs. 5c and 16). As expected, linkage between SNPs decayed with increasing genomic distance (Fig. 2d–f), although species' decay rates differed (Fig. 2e,f). Altogether, these differences in genetic diversity and structure across species probably reflect variation in recombination rates and/or the number and relatedness of sequenced genomes. The database of tag SNPs captures most within-species variation and requires up to three times less RAM compared to the full database (Supplementary Fig. 11). For the following analyses, we used the full GT-Pro database except where otherwise specified.

Optimized *k*-mer exact matching accelerates metagenotyping 100-fold. To search for exact matches between billions of sck-mers in the GT-Pro database and billions of *k*-mers in metagenome reads, it is crucial to have a search algorithm with low RAM and I/O requirements. We therefore developed an exact-match algorithm that uses data structures optimized for this specific application (Supplementary Fig. 17). After generating all *k*-mers in each metagenomic sequencing read, GT-Pro uses a *M*-bit prefix filter (*M*-filter) on the first $M < k$ bits of each *k*-mer to quickly rule out most read *k*-mers that have no chance to match database sck-mers because they do not share an *M*-bit prefix. For the *k*-mers that pass through the *M*-filter, the algorithm recruits an *L*-bit index (*L*-index; last *L* bits/suffix of encoded *k*-mer) to serve as a secondary filter that locates a bucket of presorted sck-mers in the database containing all possible exact matches to the full *k*-mer. Finally, the algorithm invokes a sequential search for exact matches between the full *k*-mer and only the sck-mers in this bucket.

We next evaluated the effect of the values *L* and *M* on GT-Pro's computational performance in both server (24 central processing unit (CPU) cores and 384 GB of RAM) and high-end laptop (eight CPU cores and 32 GB of RAM) environments. First, we measured speed and peak RAM use while tuning the values of *M* and *L*, two parameters derived from the *M*-bit filter and *L*-bit index that are expected to have a large impact on performance due to their direct relationships with query speed and RAM use. With $k = 31$ corresponding to 62 bits, we explored *M* and *L* in the range of 1–61 bits. In general, both performance metrics increase with higher values of *M* and *L* (Fig. 3a). Within the range of the tested parameters, we found best speed and peak RAM use with $M = 35$ and $L = 30$ in the laptop environment (26.5 GB of RAM) and with $M = 36$ and $L = 32$ on a server (56.55 GB of RAM). In a boundary case ($M = 36$ and $L = 30$) on the laptop where the peak RAM use hit the hardware limit, speed dropped >87%. These results demonstrate that the values of *M* and *L* should be carefully chosen based on the hardware for optimal performance. We wrote code to perform this optimization automatically when GT-Pro runs.

We then compared the computational performance of GT-Pro with tuned values of *M* and *L* to MIDAS, metaSNV[30], StrainPhlAn[31] and Kraken2[31] (Fig. 3b,e and Supplementary Table 2). Kraken2 is included to contrast speed, since it is known as a fast tool, although it does not perform genotyping. We selected 40 random

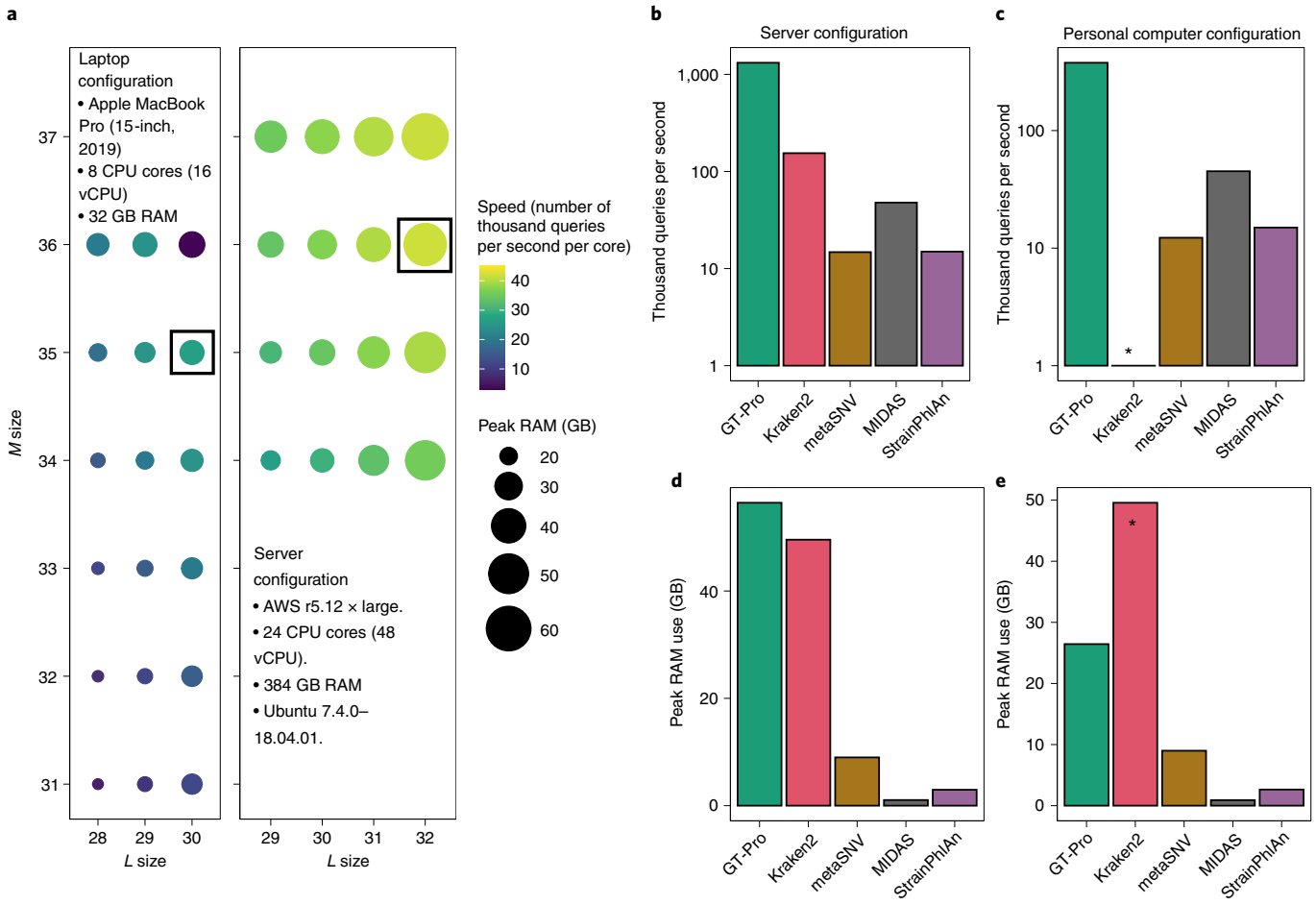

Fig. 3 | Computational performance evaluation of GT-Pro. a, Computational performance of GT-Pro in laptop (left) and server (right) environments across values for *M* (*M*-filter prefix size parameter) and *L* (*L*-index suffix size parameter), measured in bits. Since $k = 31$ corresponds to 62 bits, *M* and *L* can be in the range of 1–61. Color gradient, processing speed; circle size, peak RAM use; black box, optimal *M* and *L* for each computing environment. **b**, Server environment comparison of speed between GT-Pro, alignment-based metagenotyping (metaSNV, MIDAS and StrainPhlAn) and Kraken2. Where possible, we ran each method with its default settings (Methods). **c**, Personal computer environment comparison of speed. **d**, Server environment comparison of peak RAM usage. Peak RAM usage exceeds RAM needed to store the database due to intermediate calculations, such as applying filters. **e**, Personal computer environment comparison of peak RAM usage. These methods metagenotype different numbers of species, ranging from two (MIDAS) to 1,753 (metaSNV), due to differences in database size and species selection criteria (Methods and Supplementary Table 2). * Kraken2 does not run on the personal computer environment due to exceeding available RAM.

stool metagenomes (roughly 199 million reads) from a Tanzanian cohort[32] (Supplementary Table 9) for evaluation. Analyses were run with server (32 CPU cores and 512 GB of RAM) and personal computer (eight CPU cores and 32 GB of RAM) configurations. For each method, metagenomes were processed with their default database and settings as well as a custom database containing the same number of species ($n = 881$) as GT-Pro. These analyses showed that GT-Pro is roughly 8.5–570 times faster on a server compared with other methods (Figs. 3b) and 8.3–163.6 times faster on a laptop (Fig. 3c). On average, it took only <4 s on the server and roughly 13 s on the laptop to process each metagenome (mean, 4.97 million reads). While GT-Pro is faster than other methods, it required 1.1–53.7 times more RAM on the server (Figs. 3d) and 2.9–29.2 times more RAM on the laptop (Fig. 3e; excluding Kraken2 that ran out of memory). We attribute GT-Pro's rapid throughput to a number of factors, including an efficient search algorithm, compact data structure and optimized multi-threading. We conclude that GT-Pro data structures and algorithms greatly accelerate metagenotyping, as long as one has a computer with sufficient RAM.

Accurate identification of SNPs from simulated metagenomes. We next evaluated the metagenotyping accuracy of GT-Pro compared to MIDAS and metaSNV by running all three tools on simulated Illumina metagenomes (roughly 26 million reads) from 232 human gut isolates[33] not used to develop these methods, each with sequencing coverages ranging from 0.001 to 15 times (Supplementary Table 3). For comparability, MIDAS and metaSNV were run using a custom database of the same 881 genomes as GT-Pro; both tools implement similar approaches but use different aligners and parameters for filtering alignments. For each method, metagenotypes were compared to a set of gold standard genotypes determined from whole-genome alignment. Incorrect genotypes (false positives) result from sequencing errors and reads mapping to the wrong site, and missing genotypes (false negatives) occur when no reads map. Across metagenomes, the false discovery rate (FDR) without a vertical coverage filter was on average lowest for GT-Pro (median, 0.4%) and highest for metaSNV (median, 14.5%) (Fig. 4a, Supplementary Fig. 18a and Supplementary Table 3). FDRs were lower and differences between methods smaller when we restricted the analysis to sites in the GT-Pro database (Supplementary Fig. 20),

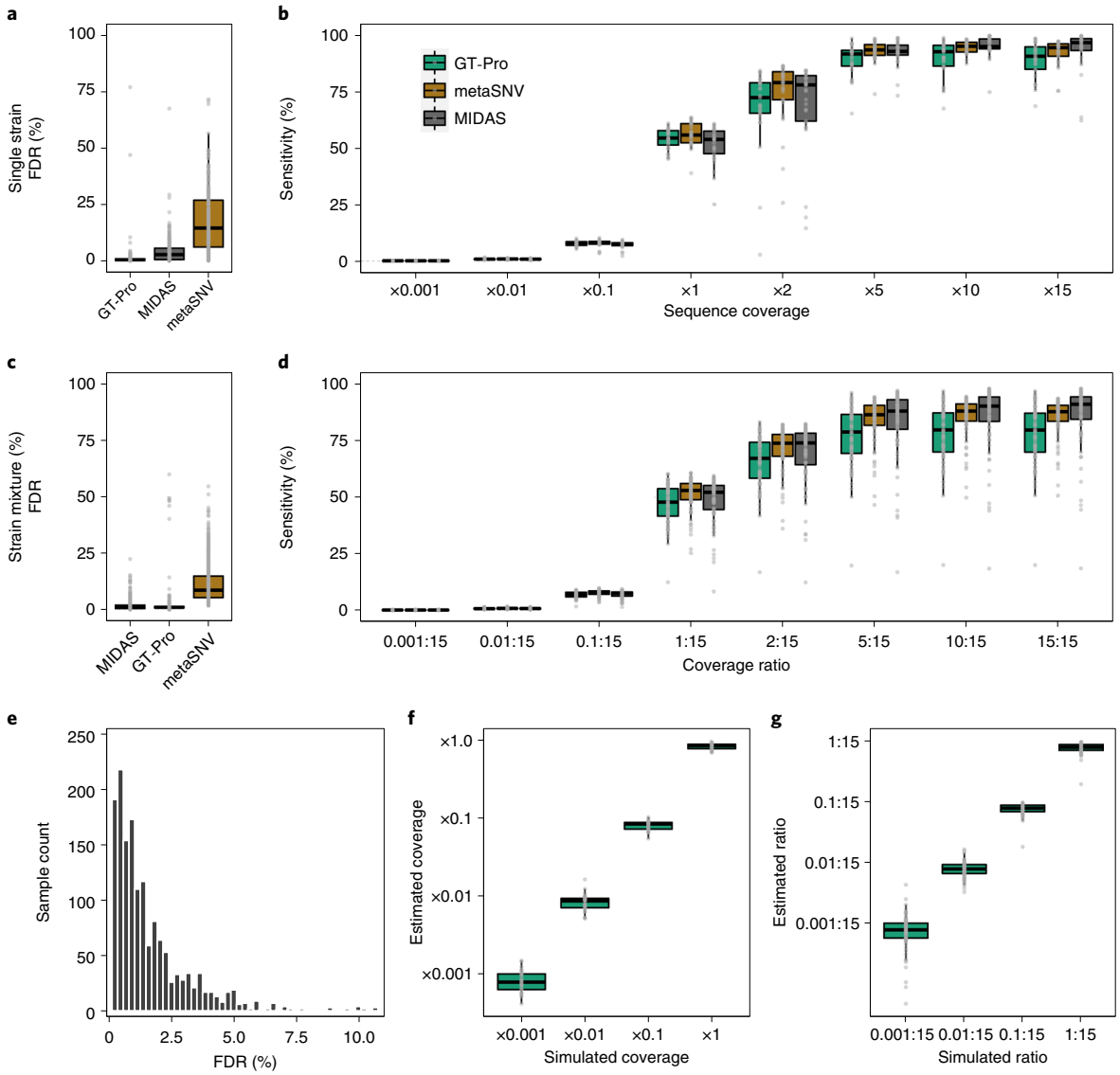

Fig. 4 | Metagenotyping accuracy evaluation of GT-Pro using simulations. Comparisons of GT-Pro, MIDAS and metaSNV metagenotyping accuracy across species based on reads simulated from isolate genomes with sequencing error. **a**, FDR comparison. Each observation is the result from a metagenome containing reads from one isolate and one value of sequencing coverage of that species ranging from 0.001 to 15 times. False discoveries are genotype calls that do not match the isolate genome from which reads were simulated (reference and nonreference alleles). **b**, Sensitivity across coverage levels from the simulations in **a** and restricted to sites in the GT-Pro database for direct comparison across methods. Sensitivity is the probability of detecting genotypes present in the isolate genome (reference and nonreference alleles) at sites in the GT-Pro database. **c**, FDR in metagenomes containing reads from two isolates of each species. A combination of sequencing coverage ratio between two isolates was simulated by fixing a more abundant isolate at 15 times coverage in all simulations, and varying the other isolate's coverage from 0.001 to 15 times (coverage ratio, 0.001:15 to 15:15). FDRs include both homozygous and heterozygous sites. **d**, Sensitivity in metagenomes from **c**. Sensitivity is the probability of correctly calling the genotype of the genomes from which reads were simulated (homozygous and heterozygous sites). **e**, FDR of genotypes imputed from tag SNPs based on allele matching in simulations in **a**. Simple imputation was done by selecting the genotype associated with the observed tag SNP. **f**, Sequencing coverage estimated using counts of reads that hit one or more sck-mers at GT-Pro genotyped SNPs correlates with the simulated coverage, even when coverage is <1×. Each observation is the estimate from metagenomic reads simulated with sequencing error from a single-isolate genome. **g**, Sequencing coverage ratio estimates based on read counts for each allele at GT-Pro genotyped heterozygous sites correlate with the simulated ratios of two isolate genomes, even when one is much less abundant than the other (≤1:15). The more abundant isolate is at 15 times coverage in all simulations.

confirming that these known SNPs are reliable genotyping targets. For the alignment methods, genome-wide FDRs were much higher at sites not matching versus matching the reference genome (Supplementary Figs. 18c and 19a). A five times vertical coverage filter brings their FDR distributions closer to GT-Pro's (Supplementary Fig. 21a) although FDRs remain high at nonreference sites that pass the filter (Supplementary Fig. 21b).

The lower FDRs of GT-Pro come with a small loss in sensitivity compared to metaSNV and MIDAS at GT-Pro sites (Fig. 4b, Supplementary Fig. 18b and Supplementary Table 4) and fewer SNPs genotyped genome-wide (Supplementary Fig. 22). Using a five times vertical coverage threshold to control the FDRs of MIDAS and metaSNV decreases their sensitivity, with no SNPs genotyped below twice the simulated coverage (Supplementary Table 4). At nonreference

alleles, GT-Pro is the most sensitive method (Supplementary Figs. 18d and 19b and Supplementary Table 4), reflecting challenges with read alignment in diverged regions that are mitigated by using 31-mer exact matching.

We next sought to evaluate metagenotyping performance in more complex scenarios. Adding unknown species to the simulation increased the FDRs of GT-Pro only slightly to a median value of 0.5% (Supplementary Fig. 23). For genotypes imputed from tag SNPs, FDRs were higher but still below 5% for most species (Fig. 4e). SNPs belonging to an LD block were five times more likely to be detected (nonzero read count) when their tag SNPs were also detected than when they were not (Supplementary Fig. 24). Finally, in simulations with all 232 species combined, GT-Pro was the least likely method to use reads to genotype the wrong species, with a lower misassignment rate (median 0.5%) compared with both metaSNV (median 22.4%) and MIDAS (median 1.8%) (Supplementary Fig. 25).

To examine how well the metagenotyping methods perform on strain mixtures, we simulated metagenomes (roughly 72 million reads) containing pairs of conspecific isolates where one strain is always at 15 times coverage and the other varies from 0.001 to 15 times coverage (Supplementary Table 5). Overall, the FDR of GT-Pro (median, 1%) was similar to MIDAS (median 0.5% at GT-Pro sites, 0.9% genome-wide) but lower than metaSNV (median 3.5% at GT-Pro sites, 8.6% genome-wide) (Fig. 4c and Supplementary Table 5). At heterozygous sites, GT-Pro's FDR (median 1.3%) was much lower than that of both MIDAS (median 20%, interquartile range (IQR) = 13.8%) and metaSNV (median 68.4%, IQR = 19.6%) (Supplementary Fig. 19c and Supplementary Table 5). GT-Pro has lower sensitivity than the alignment methods overall (Fig. 4d and Supplementary Table 6), although it is consistently more sensitive at heterozygous sites (Supplementary Fig. 19d and Supplementary Table 6). We note that alignment-based metagenotyping is typically performed with coverage and allele frequency filters to mitigate false heterozygosity calls. Applying a five times vertical coverage filter to the alignment methods improves FDR slightly (Supplementary Fig. 27), but filtering reduces sensitivity. The FDRs of alignment methods were much lower at sites in the GT-Pro database (Supplementary Fig. 26), establishing these SNPs as reliable targets for differentiating strain mixtures.

To explore whether GT-Pro can be used to quantitatively estimate species abundance, we compared the average number of sck-mer matching reads to the known genome coverage using metagenomes we simulated from individual isolates and pairs of conspecific isolates. GT-Pro slightly underestimates true coverage of each species by 12.2–23.1% (Fig. 4f) and the ratio between two strains by 20.1–22% (Fig. 4g), probably due to genomic variation and repetitive sequences causing unmapped sequencing reads. However, GT-Pro outperforms MIDAS and metaSNV across most coverage values (Supplementary Fig. 28). These results indicate that GT-Pro allele calls and counts could be used to estimate relative abundances of species and strains using a small correction factor.

Altogether, these simulations show that metaSNV and MIDAS perform well for abundant species (>5× coverage) and conserved loci, but GT-Pro is more accurate and notably more sensitive for typical coverage values and at nonreference and heterozygous sites, while being much more robust to misassignment and sequencing errors. However, in contrast to metaSNV and MIDAS, GT-Pro cannot detect novel SNPs absent from its database. We conclude that careful selection of sck-mers in conserved genomic regions enables GT-Pro to perform sensitive and specific genotyping of known SNPs from shotgun data.

Accurate metagenotyping from gut metagenomes. We next used GT-Pro to metagenotype publicly available gut microbiome samples[19,32,34,35] (Supplementary Tables 7–12) and compared results to MIDAS, metaSNV and StrainPhlAn using their default settings and reference databases. We found that GT-Pro genotyped more species (median 134) and SNPs per species with the same horizontal coverage cutoff (Supplementary Figs. 29 and 30), despite being limited to species with ≥10 genomes and unable to detect novel SNPs. This is likely due to GT-Pro having better sensitivity for low coverage species and using a human gut-focused database. For species genotyped by multiple methods, within-sample allele frequency (Supplementary Fig. 31) and across-sample allele presence and frequency (Fig. 5a–d) were highly correlated. For high-coverage species, MIDAS and metaSNV detected SNPs absent from the GT-Pro database, whereas GT-Pro detected more sites as polymorphic in medium and low coverage species (Fig. 5a–d). Despite these differences in genotyped sites, GT-Pro, MIDAS, metaSNV and StrainPhlAn produce similar estimates of pairwise genetic distances between samples (Supplementary Fig. 32), probably because rare variants missed by GT-Pro but with sufficient coverage to be genotyped with alignment-based methods represent a small fraction of overall genetic diversity. Altogether these results are consistent with our simulations and underscore the increased sensitivity of GT-Pro for species with typical sequencing coverage.

Imputing structural variation with GT-Pro SNPs. Next, we sought to determine whether GT-Pro SNPs could be used to infer the presence of nearby genes or operons, thereby serving as biomarkers for structural variants. As a case study, we used GT-Pro SNPs in flanking regions of the toxicity controlling loci CdtLoc and PaLoc in *Clostridium difficile*. Using available *C. difficile* reference genomes, we trained a random forest classifier to accurately predict the presence/absence of the toxicity loci from alleles at these SNPs (Methods and Supplementary Figs. 33 and 34). To assess the distribution of these toxicity loci in the human population, we applied the random forest models to GT-Pro metagenotypes computed for 7,459 stool metagenomes from previously published studies (Supplementary Table 13). This revealed just 34 individuals where both *C. difficile* toxicity loci were predicted (probability >0.6). These predictions were confirmed after performing whole metagenome alignments to the *C. difficile* reference genome (Fig. 5e,f). Agreement was highest in samples where *C. difficile* was abundant and when all of the genes in the locus were sequenced (Supplementary Fig. 35a,b). Several of our predictions for CdtLoc were associated with the phenotypes of hosts ($P < 0.001$), including five Crohn's disease patients who were both *C. difficile* positive and CdtLoc(+), which is consistent with high susceptibility to *C. difficile* pathology in this population[36]. In contrast, the CdtLoc locus was not predicted in most healthy infants with detectable *C. difficile*, which is consistent with common asymptomatic colonization by *C. difficile* during infancy[37]. One predicted CdtLoc(−) sample (National Center for Biotechnology Information (NCBI) accession no. SRR2566051) appeared to be a prediction error, because the patient was symptomatic of *C. difficile* infection[38]. However, *C. difficile* toxin not detected by enzyme immunoassay for the individual. These results demonstrate that GT-Pro metagenotypes can predict linked structural variants with clinical relevance. Future work is needed to determine how well this strategy generalizes to other loci and species.

Capturing novel intraspecies genetic structure with GT-Pro. Even though GT-Pro only metagenotypes known SNPs identified from reference genomes, we hypothesized that it could still detect novel strain variation based on distinct combinations of SNP alleles. To test this hypothesis, we used GT-Pro to genotype species found in 220 stool metagenomes from a recently published North American inflammatory bowel disease (IBD) cohort[35] (Supplementary Table 12). For each species detected, novel subspecies were identified by comparing the GT-Pro stool metagenotypes to genotypes of reference genomes in the database. For most species, the stool samples clustered with the reference genomes (Fig. 6a,b), suggesting that

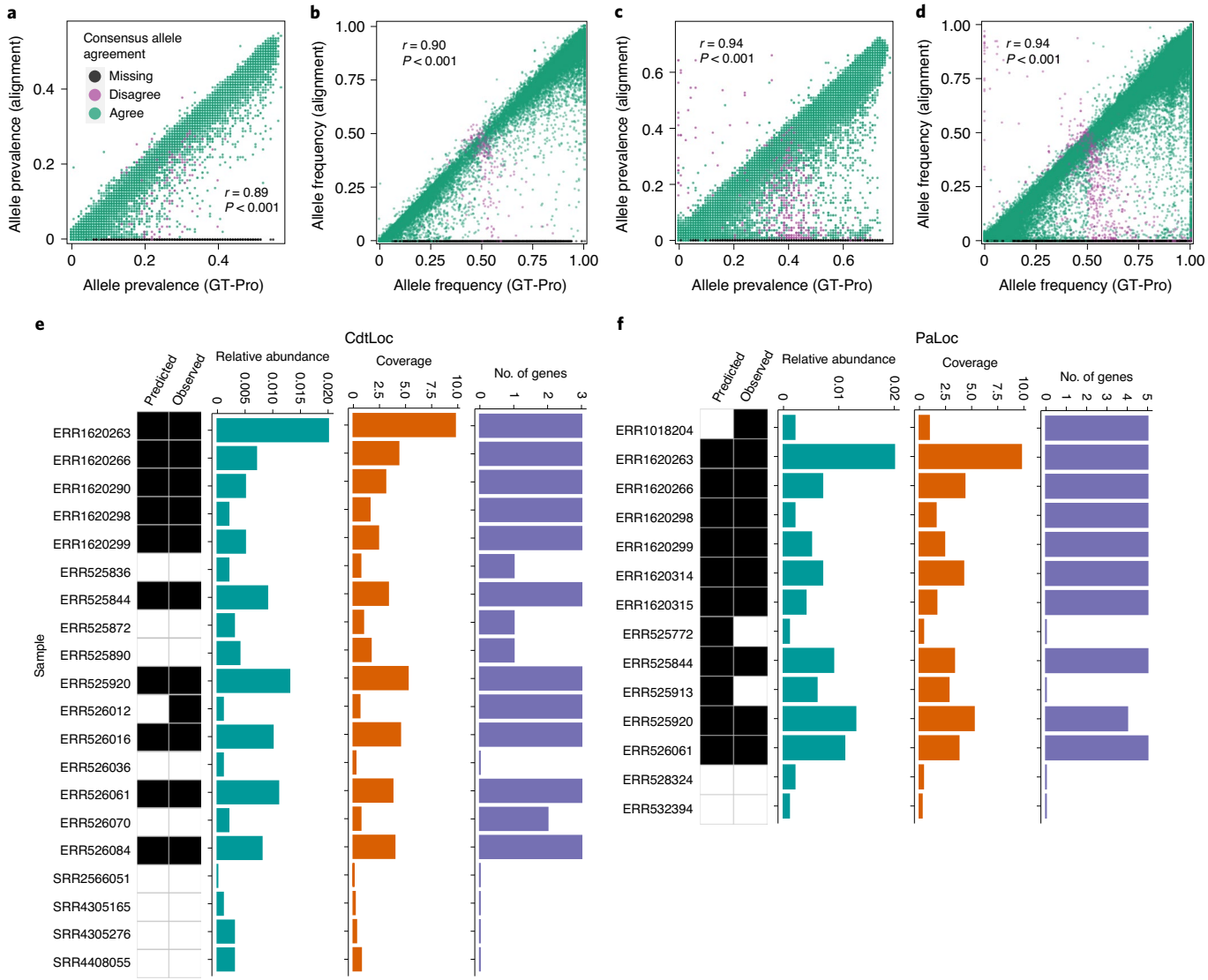

Fig. 5 | Metagenotyping and gene imputation from gut metagenomes. a, Comparison of metagenotypes from GT-Pro and MIDAS (alignment) with gut microbiome samples from a North American cohort[34] (HMP project, $n = 358$; Supplementary Table 10) for an example species, *Bacteroides stercoris* (species ID 103681). The proportion of samples in which each SNP is genotyped (prevalence) is similar for both methods, with GT-Pro genotyping slightly more samples. Each point represents a SNP, with color indicating if the consensus allele (that is, most common across samples) is the same for both methods (green), both methods return genotypes for some samples but the consensus alleles are different (purple) or only GT-Pro returns genotypes (black). **b**, Average allele frequency across samples varies across SNPs but is highly correlated (Pearson's *r*) between the two methods. Disagreements largely occur near 0.5 allele frequency, where small differences in read counts per allele can 'flip' the major and minor alleles. **c**,**d**, Comparison similar to **a** and **b** showing the species GCA_000431835.1 (genus *Succinivibrio*, species ID 100412) from a Madagascar cohort[19] ($n = 112$, Supplementary Table 11). Prevalence (**c**) and frequency (**d**). **e**,**f**, Prediction of presence/absence of *C. difficile* toxin gene loci in human gut metagenomes from a mix of cohorts ($n = 7{,}459$) (Supplementary Table 13) using a random forest classifier built with GT-Pro SNPs from flanking regions in 117 *C. difficile* isolates (Supplementary Figs. 33 and 34) with tenfold cross validation. CdtLoc genes (**e**) and PaLoc genes (**f**). Heatmaps show the predicted (first column) and observed (based on alignment, second column) presence (black) or absence (white) in each sample with detectable *C. difficile* (rows). Barplots show *C. difficile* relative abundance (left), whole-genome sequence coverage (middle) and number of detected genes from the toxin locus (right), all estimated by mapping reads from each sample to a *C. difficile* representative genome. Random Forest predictions correlate (Pearson's *r*) with abundance, coverage, and number of detected pathogenic genes (Supplementary Fig. 35).

GT-Pro's database represents the common strain diversity in these individuals. In contrast, for a few species we observed clusters of stool samples with genotype profiles that were distinct from any reference genome in the database (Fig. 6c,d), including some that were enriched for patients with IBD (Supplementary Fig. 36). To further confirm GT-Pro's ability to detect novel strains, we performed metagenotyping using simulated reads spiked into complex stool metagenomes (Methods). We found that intersample genetic distances using GT-Pro SNPs were consistent with those of other methods (Supplementary Fig. 37) and SNP detection did not deteriorate even when highly distinctive strains were present (Supplementary Fig. 38). Altogether, these results demonstrate that novel subspecies genetic structure can be discovered using GT-Pro common SNPs.

Exploring global human gut microbiome genetic variation with GT-Pro. We used GT-Pro metagenotypes to conduct a geographically

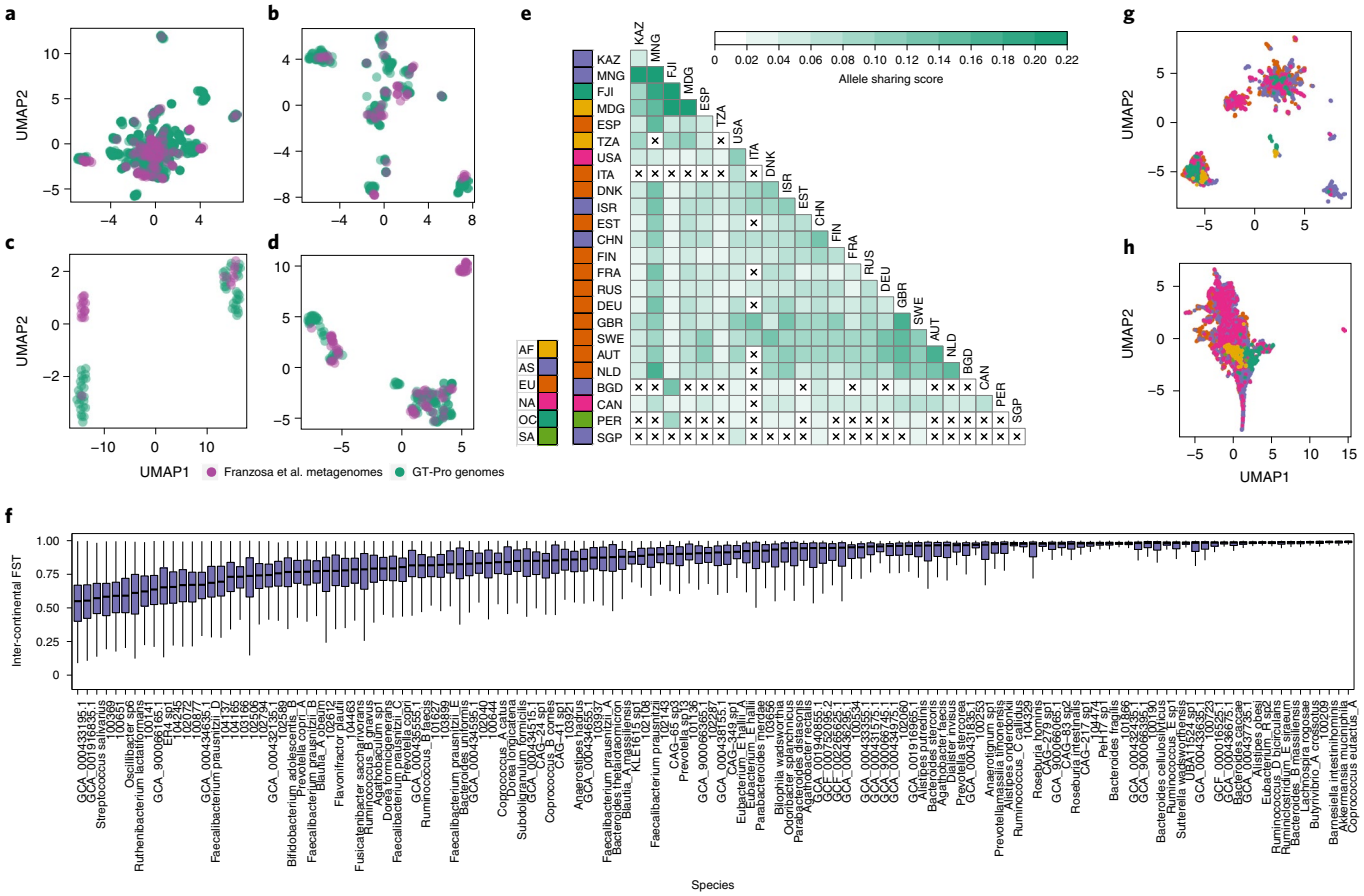

Fig. 6 | Global genetic structure in 7,459 human gut metagenomes. a–d, Gut species differ in the amount of common SNP genetic diversity already present in sequenced genomes. Metagenomic samples from a North American IBD cohort[35] ($n = 220$; Supplementary Table 12) (purple) are visualized in two dimensions alongside the UHGG genomes (green). Each plot is the result of applying UMAP to a matrix of genotypes at GT-Pro SNPs for one species. Each dot represents a strain of that species (major allele for heterozygous metagenomes); those closer together in UMAP space have more similar genotypes. **a**,**b**, *Anaerostipes hadrus* (species ID 102528) (**a**) and *Streptococcus thermophilus* (species ID 104345) (**b**) are species where metagenomes lie within the diversity previously captured by genomes. **c**,**d**, *Dorea scindens* (species ID 101303) (**c**) and *Dialister invisus* (species ID 104158) (**d**) are species where metagenomes harbor combinations of common SNPs outside the range present in genomes, which may represent novel subspecies. **e**, Heatmap of mean allele sharing scores over all species between metagenomes from different pairs of countries. Crossed cells indicate missing scores due to insufficient (<5,000) pairs of samples. AF, Africa; AS, Asia; EU, Europe; NA, North America; OC, Oceania; SA, South America; KAZ, Kazakhstan; MNG, Mongolia; FJI, Fiji; MDG, Madagascar; ESP, Spain; TZA, Tanzania; USA, United States; ITA, Italy; DNK, Denmark; ISR, Israel; EST, Estonia; CHN, China; FIN, Finland; FRA, France; RUS, Russia; DEU, Germany; GBR, United Kingdom; SWE, Sweden; AUT, Austria; NLD, the Netherlands; BGD, Bangladesh; CAN, Canada; PER, Peru; SGP, Singapore. **f**, Analysis of intercontinental population differentiation (within continent versus between continent genetic similarity, measured with the F-statistic measuring the proportion of total genetic variation captured in a subpopulation (FST)) for 78 prevalent species. Each boxplot represents a distribution of intercontinental FST for one species, ordered by medians. **g**, An example of geographic patterns captured by within-species genetic variation in the GT-Pro metagenotypes of *Agathobacter rectalis* (species ID 102492). Each dot is a metagenomic sample, colored by continent as in **e**. Dimension reduction and visualization performed with UMAP. Nearby samples in UMAP space have similar metagenotypes; distinct groups indicate geographic clustering of *A. rectalis* strains. **h**, UMAP analysis based on the relative abundances of the 881 GT-Pro species in the same samples as **g**.

diverse intraspecies genetic variation meta-analysis, encompassing 51.8 million SNPs for 881 species found in 7,459 gut samples from 31 locations across six continents (Supplementary Table 13). Expanding on and consistent with previous studies[6,8,11], we observed much less allele sharing between hosts (median 0.03, IQR 0.05) than within a host over time (median 0.38, IQR 0.4) (Supplementary Fig. 39 and Supplementary Table 14). Intrahost allele sharing across time points ranged across species and hosts from as low as between hosts to as high as technical replicates from the same sample. Interhost allele sharing differed across countries and continents (Supplementary Fig. 40a,b), generally decreasing with geographic distance (Supplementary Figs. 40a,b and 41a,b) and varying across species (Supplementary Fig. 42). Our results also show clear associations with degree of industrialization as well as relatedness of hosts (for example, hosts within villages in Fiji share more alleles than unrelated hosts in North American cities) (Fig. 6e). To identify gut species with high levels of intercontinental population differentiation, we calculated the fixation index (FST) for 78 prevalent and well-detected species (Methods) and observed large differences in the degree of differentiation (Fig. 6f). Species with high FST show distinct clusters of hosts, some but not all of which correlate with geography (Fig. 6g), consistent with lifestyle and environment playing a role in which strains colonize a host. In contrast, hosts do not cluster as clearly based on species relative abundance (Fig. 6h), emphasizing that metagenotypes may reveal microbial ecology and microbiome–host associations missed in abundance analyses.

Discussion

GT-Pro avoids computationally costly short-read alignment and performs metagenotyping with improved accuracy, especially for low coverage species. Studies at the scale of human genome-wide association studies will be computationally feasible with GT-Pro on a server, and medium-sized cohorts can be analyzed on a personal computer. We developed a GT-Pro database comprising more than 100 million common core-genome SNPs from 909 bacterial species commonly found in the human gut and demonstrated how GT-Pro can be used for strain-level exploration of thousands of microbiomes, with roughly 25,000 metagenomes analyzed in this study.

It should be noted that our method comes with several limitations. First, the GT-Pro database does not capture all human gut microbial diversity: with additional genome sequencing we hope to expand the number of SNPs and species covered. Second, GT-Pro is analogous to a genotyping array and hence does not identify novel SNPs, which require other methods, such as alignment-based metagenotyping or single-cell genome sequencing. Third, a small number of species lack species-specific sck-mers due to the presence of highly related species in the genome collection. Alternative strategies, such as using longer *k*-mers or less common SNPs, could enable GT-Pro metagenotyping for these species. Fourth, although we were very selective in the choice of genomes and SNPs used for building GT-Pro, it is impossible to exclude all imperfections (for example, incompleteness, contaminations and species misclassification) in the genome assemblies that could contribute to false SNP calls. Finally, GT-Pro does not directly genotype structural variants, which contribute to intraspecies genetic diversity[39]. However, we did show that GT-Pro can be used to impute insertions and deletions in two *C. difficile* toxin loci. Despite these caveats, we showed that the GT-Pro framework is less prone to false discoveries and roughly as sensitive as alignment-based metagenotyping tools.

We envision several directions for future work. One natural extension of this work is to combine GT-Pro with downstream algorithms to identify clusters of SNPs that represent novel microbial strains, or SNPs that accurately tag known strains in the reference database. While this study applied the GT-Pro approach to human gut microbiome, the framework could easily be expanded to other kingdoms of life and other microbial environments. Another extension is to develop alignment-free metagenotyping for short indels and structural variants. This study barely scratches the surface in terms of interpreting microbiome genetic variation. Toward using microbiomes in precision medicine, it will be critical to comprehensively identify SNPs that are associated with disease and other traits (for example, pathogenicity, antimicrobial resistance, drug degradation). We anticipate that GT-Pro will also be useful for detecting contamination, recombination and tracking variants or strains over time, host lifestyle and geography.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41587-021-01102-3.

Received: 29 June 2020; Accepted: 20 September 2021;
Published online: 23 December 2021

References

1. Garud, N. R. & Pollard, K. S. Population genetics in the human microbiome. *Trends Genet.* **36**, 53–67 (2020).
2. Maini Rekdal, V., Bess, E. N., Bisanz, J. E., Turnbaugh, P. J. & Balskus, E. P. Discovery and inhibition of an interspecies gut bacterial pathway for Levodopa metabolism. *Science* **364**, eaau6323 (2019).
3. Zeng, Q., Liao, C., Terhune, J. & Wang, L. Impacts of florfenicol on the microbiota landscape and resistome as revealed by metagenomic analysis. *Microbiome* **7**, 155 (2019).
4. Chattopadhyay, S. et al. High frequency of hotspot mutations in core genes of *Escherichia coli* due to short-term positive selection. *Proc. Natl Acad. Sci. USA* **106**, 12412–12417 (2009).
5. Treangen, T. J., Ondov, B. D., Koren, S. & Phillippy, A. M. The Harvest suite for rapid core-genome alignment and visualization of thousands of intraspecific microbial genomes. *Genome Biol.* **15**, 524 (2014).
6. Schloissnig, S. et al. Genomic variation landscape of the human gut microbiome. *Nature* **493**, 45–50 (2013).
7. Luo, C. et al. ConStrains identifies microbial strains in metagenomic datasets. *Nat. Biotechnol.* **33**, 1045–1052 (2015).
8. Nayfach, S., Rodriguez-Mueller, B., Garud, N. & Pollard, K. S. An integrated metagenomics pipeline for strain profiling reveals novel patterns of bacterial transmission and biogeography. *Genome Res.* **26**, 1612–1625 (2016).
9. Costea, P. I. et al. metaSNV: a tool for metagenomic strain level analysis. *PLoS ONE* **12**, e0182392 (2017).
10. Quince, C. et al. DESMAN: a new tool for de novo extraction of strains from metagenomes. *Genome Biol.* **18**, 181 (2017).
11. Truong, D. T., Tett, A., Pasolli, E., Huttenhower, C. & Segata, N. Microbial strain-level population structure and genetic diversity from metagenomes. *Genome Res.* **27**, 626–638 (2017).
12. Wood, D. E. & Salzberg, S. L. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.* **15**, R46 (2014).
13. Ounit, R., Wanamaker, S., Close, T. J. & Lonardi, S. CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative *k*-mers. *BMC Genomics* **16**, 236 (2015).
14. Liu, Y., Zhang, L. Y. & Li, J. Fast detection of maximal exact matches via fixed sampling of query *K*-mers and Bloom filtering of index *k*-mers. *Bioinformatics* **35**, 4560–4567 (2019).
15. Breitwieser, F. P., Baker, D. N. & Salzberg, S. L. KrakenUniq: confident and fast metagenomics classification using unique *k*-mer counts. *Genome Biol.* **19**, 198 (2018).
16. Phillippy, A. M. et al. Comprehensive DNA signature discovery and validation. *PLoS Comput. Biol.* **3**, e98 (2007).
17. Shajii, A., Yorukoglu, D., William Yu, Y. & Berger, B. Fast genotyping of known SNPs through approximate *k*-mer matching. *Bioinforma.* **32**, i538–i544 (2016).
18. 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).
19. 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).
20. Nayfach, S., Shi, Z. J., Seshadri, R., Pollard, K. S. & Kyrpides, N. C. New insights from uncultivated genomes of the global human gut microbiome. *Nature* **568**, 505–510 (2019).
21. Almeida, A. et al. A new genomic blueprint of the human gut microbiota. *Nature* **568**, 499–504 (2019).
22. Marçais, G. et al. MUMmer4: a fast and versatile genome alignment system. *PLoS Comput. Biol.* **14**, e1005944 (2018).
23. Auton, A. et al. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
24. Smith, J. M., Smith, N. H., O'Rourke, M. & Spratt, B. G. How clonal are bacteria? *Proc. Natl Acad. Sci. USA* **90**, 4384 (1993).
25. Redfield, R. J. Do bacteria have sex? *Nat. Rev. Genet.* **2**, 634–639 (2001).
26. Lin, M. & Kussell, E. Inferring bacterial recombination rates from large-scale sequencing datasets. *Nat. Methods* **16**, 199–204 (2019).
27. Ansari, M. A. & Didelot, X. Inference of the properties of the recombination process from whole bacterial genomes. *Genetics* **196**, 253 (2014).
28. González-Torres, P., Rodríguez-Mateos, F., Antón, J. & Gabaldón, T. Impact of homologous recombination on the evolution of prokaryotic core genomes. *mBio.* **10**, e02494–18 (2019).
29. Garud, N. R., Good, B. H., Hallatschek, O. & Pollard, K. S. Evolutionary dynamics of bacteria in the gut microbiome within and across hosts. *PLoS Biol.* **17**, e3000102 (2019).
30. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
31. Leinonen, R., Sugawara, H. & Shumway, M., International Nucleotide Sequence Database Collaboration. The sequence read archive. *Nucleic Acids Res.* **39**, D19–D21 (2011).
32. Smits, S. A. et al. Seasonal cycling in the gut microbiome of the Hadza hunter-gatherers of Tanzania. *Science* **357**, 802 (2017).
33. Zou, Y. et al. 1,520 reference genomes from cultivated human gut bacteria enable functional microbiome analyses. *Nat. Biotechnol.* **37**, 179–185 (2019).
34. Turnbaugh, P. J. et al. The Human Microbiome Project. *Nature* **449**, 804–810 (2007).

35. Franzosa, E. A. et al. Gut microbiome structure and metabolic activity in inflammatory bowel disease. *Nat. Microbiol.* **4**, 293–305 (2019).
36. Issa, M., Ananthakrishnan, A. N. & Binion, D. G. *Clostridium difficile* and inflammatory bowel disease. *Inflamm. Bowel Dis.* **14**, 1432–1442 (2008).
37. Rousseau, C. et al. *Clostridium difficile* colonization in early infancy is accompanied by changes in intestinal microbiota composition. *J. Clin. Microbiol.* **49**, 858–865 (2011).
38. Vincent, C. et al. Bloom and bust: intestinal microbiota dynamics in response to hospital exposures and *Clostridium difficile* colonization or infection. *Microbiome* **4**, 12 (2016).
39. 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).

Methods

Whole-genome sequences and species. To implement the GT-Pro approach for human stool metagenomics, we performed SNP discovery using the Unified Human Gastrointestinal Genomes[40] (UHGG), a collection of 286,997 microbial genomes. All genome sequences used in the study were downloaded from the UHGG at http://ftp.ebi.ac.uk/pub/databases/metagenomics/mgnify_genomes as of September 2019. We chose the UHGG database as our main genome source for four reasons: (1) the UHGG is by far the largest collection of gut microbial whole-genome sequences, (2) microbial species in the UHGG are defined over its entire set of genomes and are well annotated both phylogenetically and functionally, (3) the genome sequences in the UHGG are originally from both isolate assemblies and MAGs and (4) the UHGG excludes genome sequences that cannot be verified as coming from the human gut. The inclusion of MAGs from diverse human populations and geographic locations is critical for capturing natural genetic variation within human gut species. Since we focused on common SNPs in prevalent genomic sites (below), sequencing and assembly errors in MAGs were not likely to get called as SNPs.

To ensure reliable SNP calling, we limited our species pool to a subset of 909 species, each with at least ten high-quality (completeness ≥90% and contamination rate[18] ≤5%) whole-genome sequences. The selected species have a median of 35 genomes. Twenty-nine species with more than 1,000 genomes were further downsized to only the top 1,000 genomes in terms of quality metrics and genome length. The resulting 112,904 high-quality genomes represent most (>76%) high-quality UHGG genomes but <20% of the total number of UHGG species. As more human gut isolate genomes and MAGs are sequenced, GT-Pro can be applied to expand species coverage.

Species representative genome identification. For each species, we performed whole-genome alignment among all pairs of conspecific genomes using MUMmer[22] (v.4.0.0beta2) with the default parameters. Unreliable and repeat-induced alignments were removed using the delta-filter program from MUMmer with options '-q -r' and the remaining alignments were then extracted using the show-coords program with default parameters. The pairwise similarity between two genomes was calculated as the total number of matched bases using the following formula:

$$\text{Overall similarity} = \sum_{i=1}^{n} L_i \times ANI_i,$$

where L is the alignment length, ANI is the average nucleotide identity and i represents the ith alignment. The centroid genome with the highest average similarity to all others was selected as the representative genome for the species. This is a different approach from the UHGG study, which picked the representative genome based on the genome quality metrics only. The resulting GT-Pro representative genomes have <20% overlap ($n = 167$) with the UHGG ones, but nonetheless have high genome quality (Supplementary Fig. 2).

Genomic variation in gut microbial species. For each species, we called SNPs by aligning each conspecific genome to the representative genome using MUMmer (same parameters and filters as above). To ensure reliable SNP calling, we removed poorly aligned regions (alignment ANI < 95% of whole-genome ANI). We then concatenated all qualified alignments and classified SNPs as sites with the following characteristics: (1) two or more nucleotides, (2) presence in ≥90% of genomes and (3) minor allele frequency, ≥1%. We classified all SNPs as bi-, tri- or quad-allelic and focused on biallelic SNPs for the subsequent method implementation and downstream analyses unless otherwise mentioned. Biallelic SNPs were classified as protein coding or noncoding, and coding SNPs were annotated as (1) synonymous or nonsynonymous, and (2) non-sense or stop codon disruptions. We calculated the genome-wide ratio of nonsynonymous to synonymous polymorphism rates (pN/pS) for each species and its expected value under the assumption that mutations occurred equally across all sites, using a codon-by-codon counting method described previously[41].

SNP LD blocks. One direct benefit of calling SNPs from high-quality genomes (versus in metagenomic reads) is the ability to accurately infer haplotypes. We scanned the reference genome of each species for clusters of SNPs in strong LD ($r^2 > 0.81$). We only calculated pairwise LD for biallelic SNPs within 10-kb sliding windows for two main reasons: (1) more efficient computation and less memory consumption, and (2) LD decays below our threshold within most 10-kb windows. We calculated LD as a squared correlation for all pairs of binary-encoded SNPs for each species using PLINK[42] (v.1.9) with the following parameters:–r2,–allow-extra-chr,–ld-window 10,000,–ld-window-kb 1 and–ld-window-r2 0.81. Next, we clustered SNPs using a single-linkage strategy that is similar to the UCLUST[43] operational taxonomic unit clustering algorithm. Briefly, we iterated through all SNPs in a window, for each SNP determining whether it should be assigned to any existing cluster based on a single-linkage strategy. Namely, if a SNP is in strong LD with one or more SNPs from a cluster, it will be assigned to that cluster. If a SNP is in strong LD with members from more than one cluster, the algorithm will treat the SNP as the link between clusters and will merge those clusters. The very first SNP or any SNP that cannot be assigned to an existing cluster starts a new cluster. Similarly, we further clustered SNPs into more granular LD blocks with a higher LD threshold of 0.9 and 0.95 for all the species. These LD blocks are included in the release package of this study but are not used for further analyses.

Extraction of sck-mers. One of the main goals of this study is to leverage this SNP catalog to rapidly genotype metagenomes using short, unique genomic regions (k-mers) as markers or probes for detecting alleles that distinguish highly similar genomes from the same species in metagenomic sequencing reads. We implemented this idea by identifying any 31-base sck-mers that cover (in any of the 31 bases) each of the SNP alleles while not exactly matching other species. We chose 31 for k because (1) 31-mers can be encoded with 64-bit integers (up to 32-mer) for computational efficiency, (2) k-mers with $k \geq 20$ are fairly unique across species[44] and 31-mers are only less unique than 32-mers, (3) we prefer an odd k to avoid palindrome k-mers for the ease of processing, (4) 98.2% of SNPs do not have another common SNP within 30 bp that reduces the need for storing extra k-mers to encode multi-SNP genotypes and (5) 31-mers are slightly more storage-efficient than 32-mers allowing us to store the sequences plus essential annotations in 64-bit vectors. For each SNP we first extracted all possible 31-mers containing the SNP site from representative genome (sck-mers for reference allele). Next, we extracted sck-mers with the alternative allele by sliding a 31-bp window across the SNP site in the genome multiple sequence alignment (MSA) and selecting the most frequent 31-mer at each position. We then retrieved the reverse complements of all sck-mers. In this way, for every SNP site there will be up to 62 sck-mers targeting the reference and up to 62 sck-mers targeting the alternative allele.

Selection of species-specific sck-mers. To increase the specificity of genotyping with sck-mers, we removed any sck-mer shared by two or more species. The idea is similar to Insignia in microbial forensics, which successfully identifies signature sequences as the basis for diagnostic assays to detect microbes in DNA samples with low false positive rates. Here we adapted and scaled it up for hundreds of species to enable accurate comprehensive genotyping in whole metagenomes. For every species we generated and pooled together all possible unique 31-mers using KMC[45] (v.3.1.1) with the whole-genome sequences and the options '-k31', '-m32', '-t12' and '-ci1'. We then searched the sck-mers per species against the 31-mer pool from every other species. The sck-mers with hits against other species were excluded, which are a large fraction of sck-mers, especially for species with close relatives in the GT-Pro database. We required sck-mers for both alleles of a SNP or else the SNP was discarded.

Identification of tag SNPs. We identified a tag SNP for each LD block. Among all SNPs covered by species-specific sck-mers, we made the tag SNP the one with the most linkages ($r^2 > 0.81$) to other SNPs in the block. We estimated signs (+ or −) of LD between a tag SNP and others with the Phi coefficient, adding a pseudocount of 1 to avoid divisions by zero.

Phylogenetic resolution analysis with 16S rRNA sequences and SNPs. We performed phylogenetic resolution analysis to assess how much within-species genetic variation was captured by different subsets of genomic sites. We compared whole-genome ANI (gold standard) to cophenetic distance based on 16S rRNA genes, 40 universal single-copy phylogenetic marker genes[46], all GT-Pro SNPs and GT-Pro tag SNPs. We predicted 16S rRNA genes for all genomes using Barrnap (https://github.com/tseemann/barrnap) (v.0.9) with the option '-k bac'. For the genomes with multiple predicted copies of 16S, we selected the highest frequency gene, or the longest ones to break ties. We predicted coding gene sequences for all genomes using Prodigal[47] (v.2.6.3). We then searched these sequences against 40 universal single-copy phylogenetic marker genes with fetchMG[48] (v.1.0, http://vm-lux.embl.de/~mende/fetchMG/about.html) with default options and kept output sequences as phylogenetic marker gene sequences. For each species, the selected 16S and phylogenetic marker gene sequences were pooled separately for MSA using Clustal Omega[49] (v.1.2.4). We did not limit the use of 16S rRNA gene sequences to only cultured isolates that are absent in many species. We concatenated the SNPs per genome into a sequence, replacing missing alleles with '.' and grouping the sequences by species to obtain an MSA equivalent. We used FastTree2 (v.2.1.10) to build phylogenetic trees from the MSAs for 16S, PhyEco and SNPs. Phylogenetic distances were extracted from these trees using the function cophenetic.phylo from the ape package (v.5.3) in R. We used Spearman rank correlation to assess the concordance between the estimated phylogenetic distances and whole-genome ANIs.

Construction and optimization of the GT-Pro database. We sought to efficiently store a total of 2.8 billion species-specific sck-mers that cover 52.8 million biallelic SNPs across 881 species. For the initial optimization, we used 64-bit integers to represent the sck-mers with 00 for 'A', 01 for 'C', 10 for 'G' and 11 for 'T' and discarded the sck-mers with wildcards (for example, 'N'). The GT-Pro database was made four times smaller by this step. But it was still difficult to handle, especially on a personal computer such as the latest mid or high-end desktop/laptop models equipped with 32 GB of RAM and eight CPU cores. Since for each SNP site there are up to 124 sck-mers (31 per allele type and sequence orientation), there is a considerable amount of redundant information from the following aspects: (1) two

adjacent sck-mers can share 30 base overlap, (2) one pair of sck-mers can share up to 30 identical bases and only differ at the SNP site and (3) one sck-mer itself can be used for inferring its reverse complementary version without information loss and vice versa. These facts guided our further optimization.

To further compress the SNP data structures, we decided not to store individual sck-mers, but instead store the 61-base-long span centered at a given allele of each SNP and containing all of its possible sck-mers (sc-spans). For every SNP we extracted two sc-spans, one for the reference allele and the other for the alternative allele. To store the sc-spans from different species, we associated additional identity information with each span, including a species identifier (six decimal digits), major/minor allele indicator (0 or 1) and genomic coordinate of the SNP (up to seven decimal digits), and organized these in a binary-encoded table. Although rare, approximately 1.8% of the sc-spans have multiple SNP sites; in one extreme case, a total of 11 SNPs were found on one sc-span. We introduced additional sc-span entries to resolve each single SNP in those cases (Supplementary Fig. 12). Then we separately stored in a second table the minimal information needed to recover every sck-mer from the corresponding sc-span: a five-bit SNP offset identifying the SNP position within the sck-mer and a 27-bit index pointing to the exact entry in the sc-span table. During the look up, the sck-mer's entire sequence can be easily recovered through the formula:

$$\mathrm{sck-mer\_seq} := \mathrm{substr}(\mathrm{sc-span}, 30 - \mathrm{snp\_offset}, 31)$$

and then the reverse complementary version can be inferred. To this end, the GT-Pro database after optimization consists of two tables: (1) a 10.6-GB table of sck-mers containing a 4-byte entry for each k-mer and (2) a 2.4-GB table of sc-spans containing a 24-byte entry for each allele. The total storage required is 13 GB, twice as compact as bzip2 compression of the original sck-mer table.

Computing requirements for building the GT-Pro database. The process of building a GT-Pro database is general. There are four main steps: (1) SNP calling, (2) sck-mer extraction, (3) species-specific sck-mer selection and (4) data compression and optimization. In this study, we applied the process to the human gut microbiome and UHGG genomes. We used an assortment of Amazon Web Services (AWS) EC2 instances, including c5.18xlarge, r5.12xlarge and x1.16xlarge, to build the GT-Pro database in this study. Database building ran for about 20 d with the following peak use of computer resources: CPU, 72 cores; RAM, 976 GB and storage 15 TB. For other applications, resource consumption will depend on the number of genomes and species. While a tradeoff of time efficiency for lower hardware configuration is possible, we do not recommend building a GT-Pro database with a similar size as this study on a personal computer. However, it is tractable to build a smaller database with fewer species (for example, 50 or fewer) and genomes (for example, maximum of 100 per species) without high-performance computing.

GT-Pro development. We implemented a k-mer exact matching algorithm for metagenotyping for several reasons. First, k-mers can uniquely identify SNP sites in metagenomic reads. Second, exact matching is highly efficient to compute since it can be done in a single operation at the bit level. Third, this generates highly reproducible metagenotyping results compared with other fast search methods, such as fuzzy match or probabilistic algorithms. Fourth, it is flexible to customize by adding or dropping k-mers.

For each input metagenomic read, GT-Pro first breaks it down into 31-mers and encodes them into 64-bit integers to enable the exact matches of these 31-mers with 31-base long sck-mers. We take advantage of two ideas. First, a M-bit prefix filter (M-filter) can be used to efficiently discard sck-mers that cannot match a query k-mer (from a metagenomic read), because they do not start with the read's length M prefix (first m bits of the k-mer, $M < 2k$). Second, if the sck-mer table is colex sorted, sck-mers that end with a suffix of length L will occupy consecutive entries. Hence, we built an L-bit index (L-index) to quickly locate all sck-mers that share a given suffix by simply pointing to the first and last entries for each suffix. Implementing these strategies, the following exact-match algorithm determines whether a metagenomic 31-mer query hits the database: (1) look up query prefix with M-filter, if found, (2) look up query suffix in the L-index, if found again, (3) examine all sck-mer entries identified by the L-index one by one and report exact matches (Supplementary Fig. 17). We selected $M = 36$ and $L = 32$ as default parameters based on the best performance in our evaluation (Fig. 3). A range of values for these parameters are supported, and the optimal choices will depend on system configuration, especially the available amount of RAM.

GT-Pro counts each read only once per SNP to reduce counting bias. GT-Pro uses a concise table-shaped format for its output, in which every row represents a biallelic SNP site. Each row has exactly eight fields: species, SNP ID, contig, contig position, allele 1, allele 2 and coverage of allele 1 and coverage of allele 2.

Comparing GT-Pro to related methods. We compared GT-Pro to several other methods in a series of simulations and data analyses designed to evaluate different aspects of computational performance and accuracy, as well as to ensure our conclusions generalize across alignment-based metagenotyping strategies. We included two alignment-based methods, metaSNV[9] and MIDAS[8], which in theory both allow metagenotyping of any site across a whole reference genome. These two methods implement a similar three-step genotyping model, including aligning short reads to reference genomes (read mapping), piling up mapped reads for counting coverage per site (pile-up) and calling SNPs from site coverage profiles (SNP calling). metaSNV is a released software package (http://metasnv.embl.de/, September 2019), which uses BWA[30] (v.0.7.17-r1188) as the aligner, SAMtools[50] mpileup (v.1.9) for pile-up and is distributed with a specific database (freeze9) with 1,753 species. MIDAS (https://github.com/snayfach/MIDAS; September 2019) uses Bowtie2 (ref. [51]) (v.2.3.2) for read mapping, PySAM (https://github.com/pysam-developers/pysam; v.0.12.0.1) for pile-up, and a default database with 5,952 species. Both methods allowed us to use the representative genomes from GT-Pro. We also used a version of MIDAS with improved parallel processing for higher metagenotyping throughput without functional difference. To avoid possible bias, we did not use this version for the computing performance evaluation.

We also included another alignment method, StrainPhlAn[11] (v.3) in the evaluations for which this tool was applicable. Although StrainPhlAn uses a similar model to metaSNV and MIDAS, it differs in several important ways. First, it maps short reads to a set of species-specific marker genes in lieu of a whole reference genome. This step should make read mapping more efficient. However, it prevents the metagenotyping of SNPs located outside the marker genes. Second, the species-specific marker genes are preselected and cannot be user modified. Thus, we cannot practically use StrainPhlAn with the GT-Pro database. Furthermore, StrainPhlAn does not output individual genotypes. Instead, it concatenates them into consensus allele sequences with stringent per-sample and across-sample filtering, and it generates a MSA and phylogenetic tree from filtered consensus alleles as its standard output. For the last two reasons, we excluded StrainPhlAn from the analyses that require comparing individual metagenotypes at specific genomic coordinates (that is, using the same reference genome).

For our evaluations of computing performance, we included Kraken2[12] (v.2.0.9-beta), a k-mer based method. Kraken2 is not a genotyping method nor does it use alignment, but it is widely endorsed for high read processing speed. We included it to provide an additional reference in terms of GT-Pro's computing efficiency.

Unless otherwise noted, we ran GT-Pro with the full database and default parameters and removed from output the sites with less than two times coverage to reduce data size (that is, the number of sites). We ran both metaSNV and MIDAS with default parameters except a minimum site coverage of five instead of ten times as suggested. StrainPhlAn has a preset of well-tuned parameters, so we ran it with default settings in every case. For all paired-end samples, we processed only forward read (fastq 1) for the simplicity of comparison and analysis. We skipped a sample if it did not have a forward read sample as extracted from a Sequence Read Archive (SRA) file using fastq-dump in SRA toolkit[31] (v.2.10.0). However, we note that using the reads of both directions can effectively increase the coverage, which thus should be recommended especially for species with low abundance.

Computing performance evaluation on publicly available metagenomes. The shotgun metagenomic samples used for performance evaluation were human gut microbiome fecal samples from Hadza hunter-gatherers in Tanzania[32] (Supplementary Table 9), which are available at https://www.ncbi.nlm.nih.gov/ with accession ID SRP110665. The selection of samples and reads was random. For all methods, the chosen 40 samples were processed in their entirety.

First, we evaluated the performance of each method with its default settings. GT-Pro and metaSNV (GT-Pro and metaSNV.default) by default genotype all species included in the database. MIDAS by default (MIDAS.default) only genotypes the most abundant species in a sample that are above its predefined coverage threshold. We found it only genotyped two species on average (median) per sample in our test (Supplementary Table 2). We ran StrainPhlAn (StrainPhlAn.default) for the species clades with nonzero relative abundance as it reported as the default. It genotyped 50 species on average (median) per sample in our test. To ensure a fair comparison between different methods, we accounted for the difference in database size when possible. We ran both metaSNV and MIDAS with the same number of species as GT-Pro database (metaSNV.881 and MIDAS.881). We limited the running of StrainPhlAn to a total of 881 species-level clades that were randomly selected (StrainPhlAn.extract_881). For further contrast, we also ran StrainPhlAn without its end step (StrainPhlAn.no_extract), which is essentially MetaPhlAn[52]. We ran GT-Pro with the full database and Kraken2 with the standard database.

We calibrated the performance of GT-Pro on two computing environments (server and laptop), as we tuned the values of M and L: two parameters that are derived from the M-filter and L-index. For server-based evaluation we set up an AWS EC2 instance with the following specifications: AWS r5.12xlarge, 24 physical CPU cores (48 vCPU), Intel 8175M CPU at 2.50 GHz and 384 GB of RAM. The laptop environment was on an Apple MacBook Pro (15-inch, 2019) with the following specifications: eight physical CPU cores (16 vCPU), Intel(R) Core(TM) i9-9980HK CPU at 2.40 GHz and 32 GB of 2,400 MHz DDR4 RAM. For the performance comparison of all methods, we used another two computing environments (server and personal computer) with tuned M and L. For server-based evaluation we set up an AWS EC2 instance with the following specifications: AWS r5.16xlarge, 32 physical CPU cores (64 vCPU), Intel(R) Xeon(R) Platinum 8259CL CPU at 2.50 GHz and 512 GB of RAM. The personal

computer environment was on AWS c5.4xlarge with the following specifications: eight physical CPU cores (16 vCPU), Intel(R) Xeon(R) Platinum 8275CL CPU at 3.00 GHz and 32 GB of RAM.

Each method was run using all cores of an environment when possible. We measured both wall clock time and peak RAM use for each method using GNU Time (v.1.7). MIDAS.881 was not able to finish normally the evaluation samples in the designated maximum running time window (24 h) regardless of environment and thus terminated manually. We estimated the speed cap of MIDAS.881 with the maximum time use of 24 h and omitted peak RAM use. Kraken2 did not start running in the personal computer environment due to the RAM limitation. We instead collected the RAM requirement of Kraken2 as the peak RAM use for the personal computer environment and omitted the estimate of speed.

Accurate identification of SNPs from simulated metagenomes. To evaluate the accuracy of metagenotyping by GT-Pro and to compare it to alignment-based metagenotyping, we conducted a series of simulations that used reads generated from whole-genome sequences. We used InSilicoSeq[53] (v.1.4.2) with the options '–model HiSeq' and '–n_reads 2000000' to simulate reads with Illumina length and error characteristics. This generated two paired-end read files each containing roughly 1 million 126-bp-long reads from each genome. For simplicity, we proceeded with only the forward reads. We varied genome coverage across simulations by randomly drawing reads from the simulated metagenomes. The number of reads required for a level of coverage c was estimated by the following formula: number of reads $= c \times$ genome length/126. For example, to provide a ten times coverage for a genome with a size of 5 megabp (Mbp), a rough number of 396,825 ($10 \times 5{,}000{,}000/126$) 126-bp-long reads is needed. We used eight levels of coverage: 0.001, 0.01, 0.1, 1, 2, 5, 10 and 15 times.

We used isolate genomes ($n = 978$) from the Culturable Genome Reference (CGR) study[33] for these simulations. These genomes were also included in the UHGG genome collection under the label BGI and covered in GT-Pro database. These genomes were cultivated from fecal samples of healthy humans and characterized as nonredundant and high-quality draft genomes. To ensure GT-Pro did not gain an advantage due to the fact that the genomes used for simulation were also included in the UHGG genome collection and contributed SNPs to GT-Pro database, we removed information from these isolates from the GT-Pro database by excluding the isolates and rebuilding the database from scratch for the purpose of these simulations.

First, we simulated a simple scenario where each metagenome only contained reads from a single isolate. We selected a total of 232 isolate genomes from 152 species with at least 5,000 SNPs covered in the GT-Pro database. Each isolate genome was randomly assigned a sequencing coverage from the eight levels. To evaluate metagenotyping accuracy, we identified SNPs using whole-genome alignment between each genome used to simulated reads and the GT-Pro reference genome for that species. We additionally imputed genotypes within LD blocks by assigning the reference allele if the tag SNP had the reference allele, and the alternative allele if the tag SNP had the alternative allele. SNPs outside LD blocks and SNPs with Phi<0 with their tag SNP were not imputed.

In a second scenario, we generated sequencing datasets containing reads from a pair of conspecific genomes to simulate metagenomic samples with strain mixtures. We used a total of 80 genome pairs, each from a distinct species. For each pair, we generated eight metagenomic samples by fixing the sequencing coverage for one genome at 15 times and generating reads for the other genome with each of the eight coverage levels. In this way, we generated a total of 640 metagenomic samples from 80 species with the coverage ratio of a strain 1 versus strain 2 ranging from 0.001:15 to 15:15. To evaluate accuracy, we called SNPs in the whole-genome alignment of the two strains used to simulate reads to each other and to the reference genome. Based on the alignments, each GT-Pro SNP was labeled as a homozygous reference allele, heterozygous or homozygous alternative allele.

We ran GT-Pro, metaSNV and MIDAS on the simulated reads from each simulation. No minimum site coverage or any other filtering criteria were applied to any methods in the primary analysis to minimize the biases that could be introduced by this preprocessing. We separately evaluated the effect of a five times vertical coverage threshold applied to the two alignment methods, as has been recommended by their developers.

Next, we compared metagenotypes to the known values and estimated various performance measures. Our evaluations focused on correct identification of SNPs across coverage values and on the accuracy of the alleles in the metagenotypes, as well as accurate estimation of species and strain abundance using SNP site coverage and our ability to use tag SNPs to impute genotypes of other GT-Pro SNPs. Since we separately estimated the genetic diversity that can be captured by GT-Pro SNPs in genomes (above), these evaluations focused on correct allele calls at GT-Pro SNP sites (that is, we only included SNPs that are potentially able to be genotyped by both GT-Pro and our bowtie-based pipeline). We separately evaluated the performance of the alignment methods MIDAS and metaSNV genome-wide, as they would ordinarily be run.

In simulations with a single species per metagenome, false positives (FP) are incorrect genotypes as well as any genotypes called for species other than the one from which reads were generated. True positives (TP) are correct genotypes in the species from which the reads were generated. Hence, we can efficiently compute the FDR as one minus the ratio between the sum of true positive sites and the sum of all reported sites. False negatives (FN) are sites with no genotype or an incorrect genotype, so sensitivity can be calculated as TP/(TP + FN) or the ratio of true positive sites to all GT-Pro sites. These statistics were calculated for each method on nonreference sites only and on all sites (reference and nonreference alleles) separately.

In simulations with strain mixtures, we estimated FDR and sensitivity for calling heterozygosity (that is, two genomes have different alleles), as well as accuracy of genotypes (which can be homozygous reference, heterozygous or homozygous alternative). These statistics were calculated for each method on heterozygous sites only and all sites (homozygous reference, heterozygous and homozygous nonreference) separately. At homozygous sites, false positives are heterozygous calls and homozygous calls with the incorrect allele.

To evaluate whether GT-Pro is overfitting to the genomes in its database, we performed a simulation that included unknown species. We used fivefold cross validation to determine the accuracy of GT-Pro on simulated metagenomes that include species not used to build the database. We randomly shuffled the species in the GT-Pro database and divided them evenly into five groups, each accounting for roughly one-fifth of all species. We then generated a total of five GT-Pro databases and for each one we held out a species group from the $n-1$ species filtering for species-specific sck-mer screening. This procedure simulates the case where each GT-Pro database has a distinctive group of unknown species. Next, we ran GT-Pro with each database on the reads simulated from the single genomes that included the unknown species. We ran GT-Pro, metaSNV and MIDAS, and for each method we calculated FDR values as in Fig. 4a.

To directly estimate the misassignment of reads to other species, we performed a simulation of a microbiome community with a high sequencing coverage, which amplifies the chance of observing any misassignments. Using the same 232 isolate genomes (152 species) as in the single-isolate simulations, we generated reads at 15× per genome. We ran MIDAS, metaSNV and GT-Pro on the simulated reads without applying any coverage cutoff. A misassignment was defined as a genotyped SNP with reads for at least one allele coming from any other species. The rate of misassignment per species was calculated as the ratio between misassignments and total genotyped SNPs.

To mitigate false positive genotyping of species that is absent in a metagenome, we also used this simulated high-coverage metagenome to find appropriate default filtering criteria. Since misassigned reads typically occur at only a subset of SNPs with local homology to another species, we reasoned that absent species with some falsely genotyped SNPs would have low horizontal coverage, which we defined as the proportion of SNPs genotyped (that is, with reads). From the output of GT-Pro, we used the simulated origin of each read to count total false positive sites as a fraction of all sites in the GT-Pro database for each species. We then compared these fractions to the horizontal coverage of the species, which is high when the species is present. We observed that a cutoff as low as 0.025 on horizontal coverage keeps the rate of falsely detecting and genotyping a species below 2% (Supplementary Fig. 29). Based on these results, we chose a more stringent cutoff of 0.05 horizontal coverage to determine the presence of a species.

Metagenotyping 25,133 human microbiomes with GT-Pro. To further demonstrate the computational efficiency of GT-Pro, we applied GT-Pro to metagenotype a total of 25,133 publicly available human metagenomes from diverse body sites, biogeographic regions, lifestyles and age groups (Supplementary Tables 7 and 8). These samples comprise >40 trillion bases in >400 billion reads from 45,683 fastq files. We used two AWS EC2 instances (type: c5.18xlarge) for processing the samples, half on each instance. Each instance has 72 vCPUs, 144 GB of RAM, 25 GB per second of network bandwidth and 19 GB per second of Elastic Block Store (EBS) bandwidth. Due to the limits of EBS storage space, we wrote a Bash script for processing the samples in small batches with four steps: (1) download SRA files with accession numbers, (2) extract fastq files from SRA files, (3) run GT-Pro on fastq files and (4) deposit metagenotyping results to S3 storage and clean out EBS space. It took less than 3 weeks to process these samples. We estimated that the same task would take more than 1 year on the same system using alignment-based metagenotyping. We observed that GT-Pro shifts the computational bottleneck from genotyping reads to data downloading, transferring and decompression, which were responsible for two-thirds of its processing time.

Accurate metagenotyping from gut metagenomes. To further evaluate GT-Pro, metaSNV, MIDAS and StrainPhlAn on complex communities, we metagenotyped gut metagenomes from three healthy cohorts (Supplementary Tables 9–11) from Tanzania[32] ($N = 40$), North America[34] (HMP; $N = 358$) and Madagascar[19] ($N = 112$), as well as a North American IBD cohort[35] (Supplementary Table 12). The IBD cohort is from a 2019 paper and did not contribute any MAGs to the UHGG collection or GT-Pro sck-mer database. We chose diverse samples to explore the applicability of GT-Pro to human gut microbiomes across populations.

With the Tanzania cohort, we compared GT-Pro to the alignment methods with default databases to evaluate how the different databases affect metagenotypes. From the output of the three methods, we counted the number of species and median number of sites genotyped per species. A horizontal coverage cutoff of 0.05 was applied to all three methods. For species detection, no vertical

coverage cutoff was applied. For median number of sites genotyped per species, we used the recommended, default vertical coverage cutoffs for each method: 5× and 10× with MIDAS and metaSNV and 1× with GT-Pro.

With the HMP and Madagascar cohorts, we compared GT-Pro to the alignment methods using GT-Pro reference genomes (that is, same genome database) to isolate the effect of *k*-mer exact matching versus alignment. We compared within-sample heterozygosity and across-sample allele presence and frequency between methods, using only GT-Pro SNP sites (that is, sites that can theoretically be genotyped by both methods, as in simulations). The methods were compared using Jaccard index (a concordance estimate) on genotyped sites and Pearson correlation of genotypes. Jaccard distance was calculated as the number of genotypes (major allele) that agreed by both methods divided by the number of SNP sites genotyped by both methods. Pearson correlation coefficient was calculated using the nonzero allele frequencies at sites genotyped by both methods. The consensus allele was identified per SNP site as the genotype with the higher across-sample allele presence. These statistics were computed using genotypes from all species per sample and separately for each species.

With the North American IBD cohort, we analyzed the consistency between GT-Pro, metaSNV, MIDAS and StrainPhlAn. We ran MIDAS and metaSNV on the samples with both 5 and 10× vertical coverage cutoffs. Both GT-Pro and StrainPhlAn were run with default parameters. We concatenated major alleles from the output and calculated pairwise genetic distances between samples using RAxML[54] (v.8.2.12). Genetic distances between samples were compared using reported genotypes per species. Here the alignment-based methods could use SNPs that are not covered by the GT-Pro database, and StrainPhlAn is limited to SNPs in its database of marker genes.

Flanking genes analysis of *C. difficile* pathogenicity locus. We conducted an analysis to test if structural variants can be predicted using flanking GT-Pro SNPs. We used the medically important example of presence/absence of genomic regions containing toxin genes in *C. difficile* strains. First, we investigated SNPs from two 5′ (CD2601 and CD2602) and one 3′ flanking gene (trpS) to predict the presence/absence of the binary toxin CDT locus (CdtLoc), which contains the binary toxin encoding gene cdtAB and a response regulator gene cdtR (Supplementary Fig. 33a). Second, we used SNPs from two genes (cdd1 and cdu1) as well as the regions in between that flank the pathogenicity locus (PaLoc; contains five toxin genes tcdA, tcdB, tcdC, tcdE, tcdR) to predict the presence/absence of this region (Supplementary Fig. 33b). In contrast to the PaLoc region, which has several different known structural variants, the CdtLoc region only has one known deletion where a unique 68-bp sequence replaces cdtAB and trpS.

For both loci, we first applied machine learning to GT-Pro SNPs in *C. difficile* genomes to determine whether the presence/absence of the toxin gene could be predicted accurately from the SNPs. If so, it would establish that GT-Pro SNPs are tightly enough associated with the toxin gene structural variants that we could apply GT-Pro to metagenomes and infer toxin gene presence.

To build a genome data set, we downloaded 1,171 *C. difficile* genomes deposited in NCBI RefSeq (https://ftp.ncbi.nlm.nih.gov/genomes/refseq/bacteria/Clostridioides_difficile/) as of May 2018. We used checkM[18] (v.1.0.11) to estimate the genome quality. We removed the genomes with completeness <0.97, contamination >0.02, strain heterogeneity >0.25, GC content >30% or genome size >4.6 Mbp. We also used dRep[55] (v.1.4.3) to remove highly similar genomes. Any genomes with ANI cutoff >0.99 and minimum coverage of the larger genome >0.95 to a nonredundant genome were removed. This resulted in 114 high-quality, nonredundant *C. difficile* genomes (Supplementary Table 15). Among these, we proceeded to analyze 110 where the flanking genes for both toxin loci were present.

We annotated the CdtLoc and PaLoc regions in the *C. difficile* genomes using BLAST[56] (v.2.7.1), requiring both >20% horizontal coverage and >90% sequence identity to call hits to each toxin gene. We found that this subset of 110 *C. difficile* strains contains a variety of deletions in the PaLoc region, which could result from natural structural variants or incomplete genome assemblies. In contrast, 15 strains harbored the known 68-bp substitution in CdtLoc and the rest had the complete locus.

Next, we investigated what GT-Pro SNPs were found in and flanking CdtLoc and PaLoc within the genomes. We found no SNPs in the CdtLoc region (confirmed by alignment of all sck-mers to the *C. difficile* genomes), probably reflecting the fact that this region is differentially present across strains and hence contains no highly prevalent SNPs. But we did identify 180 SNPs located in the flanking genes: 62 in CD2601, 60 in CD2602, 34 in trpS and 24 in intergenic regions (Supplementary Fig. 33a). For the PaLoc region, we identified 56 SNPs in the flanking regions: seven in cdd1, 32 in cdu1 and 17 in between the two genes (Supplementary Fig. 33b). We removed SNPs with the same genotype in all genomes because they have no power to differentiate genomes for the purpose of classification and kept 148 and 40 SNPs for the CdtLoc and PaLoc region, respectively.

To test whether these GT-Pro SNPs in flanking regions can be used to infer the structural variants, we built a random forest classifier for the presence/absence of CdtLoc (and separately for PaLoc) using flanking SNPs as features. We labeled a genome toxigenic only if it has a complete PaLoc region (for example, presence of five genes) and the partial presence of a gene as presence. To ensure classifier to correctly deal with genomes or samples with all flanking SNPs missing or absent, we added to training a mock genome with all feature SNPs and prediction label marked as absent. The random forest classifier was then trained with randomForest package (v.4.6-14) in R with default hyperparameters and run through tenfold cross validation. Predictive power of each SNP was determined using importance function in the same package, which quantifies the contribution of each variable to prediction accuracy by permuting it.

After establishing that GT-Pro SNP allele values were associated with flanking toxin gene presence/absence, we proceeded to metagenotype these SNPs in human stool metagenomes with GT-Pro to predict which ones harbored *C. difficile* with CdtLoc and/or PaLoc. We screened genotype profiles produced by GT-Pro in a total of 7,459 gut samples. We found 135 and 52 samples containing one or more flanking SNPs for the CdtLoc and PaLoc region, respectively. For the simplicity of analysis, we only continued on a subset of samples for both the CdtLoc ($n=41$) and PaLoc ($n=33$) region, which contained all top 20 SNPs with highest feature importance. We extracted and binarized the flanking SNPs as feature inputs and classified samples with the aforementioned random forest classifier trained on genomes. The predicted presence/absence of the PaLoc or CdtLoc region in a sample was determined by probability >0.6.

For the same samples, we used an alignment method (MIDAS species and snps pipelines) to determine observed presence/absence of the CdtLoc and PaLoc region, *C. difficile* relative abundance, whole-genome sequence coverage, and number of detected genes from the pathogeneticity locus. Reads from each sample were mapped to a *C. difficile* representative genome (UHGG ID GUT_GENOME142303). The presence of a gene was determined by horizontal coverage (>50% of full sequence length) by reads, and the observed presence of a region was determined by percentage of genes present (≥80%), that is presence of the PaLoc region requires at least four genes present and CdtLoc requires all three genes present.

Novel subspecies detection in IBD microbiomes. To explore genetic structure within species, we selected a set of metagenomic samples ($n=220$) from a North American IBD cohort[35]. Since these data are not related to the GT-Pro database, the strains in the metagenomes are likely to be at least novel compared to the genomes we used to build the GT-Pro sck-mer database. We sought to determine whether the common SNPs covered by the GT-Pro database could be used to identify novel allele combinations, which could arise through homologous recombination, horizontal transfer or recurrent mutation. To do so, we applied GT-Pro to the 220 metagenomes and then performed dimension reduction on the resulting metagenotypes along with the genotypes of the genomes used to build the sck-mer database. If metagenome-specific samples cluster outside the variation of genomes in the dimension reduction, it would suggest that they carry novel SNP allele combinations. A cluster of metagenomes would indicate a group of related strains that differ from known strains and could represent a new subspecies.

We used the following steps for performing this exploratory analysis: (1) identified and estimated the frequency for the major allele per SNP per sample, (2) identified the major allele per SNP as the allele present in the most samples, (3) recorded the matrix of major allele frequency for each sample, (4) extracted a similar matrix of binary major allele presence/absence for each genome used to build the GT-Pro sck-mer database and (5) applied uniform manifold approximation and projection (UMAP) analysis using umap package (v.0.2.3.1) in R and plotted metagenomes and genomes in UMAP coordinates. To account for possible bias due to fewer covered SNPs, here we applied two main filters to both genomes and metagenomes: (1) we excluded a SNP in both genomes and metagenomes from this analysis if it is present in fewer than five metagenomes and (2) we excluded a metagenome if it has <1,000 genotyped SNPs or <50% of total SNPs genotyped. In the four example species shown in Fig. 6, most SNPs (72–84%) were genotyped in the metagenomes on average. To further check whether the metagenomes in Fig. 6c (*Dorea scindens*; species ID 101303) and Fig. 6d (*Dialister invisus*; species ID 104158), which showed novel subspecies, were affected by missing SNPs, we compared the number of SNPs genotyped in those with all metagenomes using Welch two-sample *t*-test. For both species, we do not find a significant difference ($P=0.44$ and 0.49 for *Dialister invisus* and *Dorea scindens*, respectively), which suggests the novel subspecies delineated in those metagenomes was unlikely due to a disproportionate number of missing SNPs.

To further evaluate the capability of GT-Pro to detect novel subspecies, we genotyped metagenomes with simulated reads of unknown strains spiked in and compared GT-Pro and alignment methods using the resulting estimates of genetic distances between strains. For this analysis, we selected a total of 247 genomes from an arbitrary species (ID 100113), which included four CGR genomes. There were no CGR genomes in the GT-Pro database we used for this accuracy evaluation, making them unknown to GT-Pro. We simulated reads at 15× coverage per genome and spiked each read set into one arbitrary sample (accession no. SRR6468562) from the Franzosa et al. cohort. We ran MIDAS, StrainPhlAn, metaSNV and GT-Pro on the resulting samples. We applied a five times coverage cutoff to both MIDAS and metaSNV and no coverage cutoff to GT-Pro. We concatenated major alleles from the output of each sample and calculated pairwise genetic distances between samples using RAxML.

Finally, we sought to investigate whether GT-Pro recovers disproportionately fewer SNPs from highly distinctive strains, since it is based on common SNPs. We

selected a total of 24 genomes (species ID 102295) from the GT-Pro database with between 1,812 and 40,257 SNPs compared to their reference genome (aligned using MUMmer). We then simulated reads at a high coverage (15×) from each of these genomes and spiked them into an arbitrary sample (accession no. SRR6468551). We ran metaSNV, MIDAS and GT-Pro on these spiked samples and left out StrainPhlAn for aforementioned reasons (noninterchangeable reference genomes and output limitations). We counted SNPs recovered by each method without applying any coverage cutoff to be as comparable as possible. As expected, GT-Pro detected fewer SNPs compared to metaSNV and MIDAS, since it by design covers a fraction of total SNPs. However, we did not find the rate of SNP recovery to deteriorate with increasing distance to the reference genome. This held when distance was computed using only the SNPs in the GT-Pro database (common, with species-specific *k*-mers) and when using all SNPs. In the latter case, GT-Pro had a slightly lower correlation with distance to the reference genome compared to metaSNV and MIDAS. The highest correlation was for GT-Pro when genetic distance was computed using only SNPs in the GT-Pro database, suggesting that GT-Pro's lower overall SNP recovery rate was mostly not due to failure to recover common SNPs.

Global population genetic diversity of human gut microbiomes. From all metagenotyped samples, we selected a total of 7,459 gut samples from 31 locations across all continents except Antarctica (Supplementary Table 13) for a global population genetic diversity analysis. The selection is limited to human gut samples and a single sample per host to minimize potential biases that repeated sampling might introduce.

We calculated an allele sharing score between each pair of samples as an estimate of their genetic similarity. We identified shared SNP sites as those genotyped in both samples, and then defined the allele sharing score as the number of shared SNP sites with the same major allele in both samples divided by the number of all genotyped SNP sites in the lesser sample, that is, min(no. of genotyped sites in sample 1, no. of genotyped sites in sample 2). We adopted this measure of similarity because many samples are very different in the number of reads.

We down-sampled 84,000 pairs of samples and calculated the FST statistic, an estimate of population differentiation, for 78 prevalent and well-detected species (present in >500 intercontinental sample pairs, each pair has >1,000 shared SNP sites). To reduce bias due to disparate sample size, we performed down-sampling by randomly drawing 6,000 pairs of samples per continent pair and ignored continent pairs with <6,000 pairs of samples. FST is defined as 1 minus the ratio of within-sample and between-sample nucleotide diversity (π), or more formally, $1-(\pi 1/2+\pi 2/2)/\pi b$, where πi ($i=1$ or 2) represents within-sample π for sample i and πb represents the between-sample π between samples 1 and 2. To estimate within-sample π, we first calculate the single site π of a genotyped SNP site as $2\times a1\times a2/(a1+a2)/(a1+a2-1)$, where ai ($i=1$ or 2) represents the coverage of the ith genotype of a SNP site, and then take an average of single site π across all genotyped sites. Similarly, to estimate between-sample π, we first calculate the single site π of a shared site as $(a11\times a12+a21\times a22)/(a11+a21)/(a12+a22)$, where aij (i, $j=1$ or 2) denotes the coverage count of the i'th genotype of a SNP site in sample j, and then take an average of single site π across all shared sites. This approach only accounts for biallelic SNPs.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

All described datasets are publicly available through the corresponding repositories. Genome assemblies for building GT-Pro used in this study were downloaded from the UHGG database and are available at MGnify (http://ftp.ebi.ac.uk/pub/databases/metagenomics/mgnify_genomes). The 1,171 *C. difficile* genomes are available at NCBI RefSeq (https://ftp.ncbi.nlm.nih.gov/genomes/refseq/bacteria/Clostridioides_difficile/), and the accession numbers of 114 high-quality nonredundant *C. difficile* genomes are in Supplementary Table 15. All metagenomic samples are available at NCBI SRA (https://www.ncbi.nlm.nih.gov/sra) with accession numbers in supplementary tables: 25,133 human microbiome samples (Supplementary Table 8), Tanzania (Supplementary Table 9), North America (Supplementary Table 10), Madagascar (Supplementary Table 11) and North American IBD cohort (Supplementary Table 12) and global biogeography samples (Supplementary Table 13). The GT-Pro SNP databases and genotype profiles of 25,133 human microbiome samples generated in this study are available in a cloud server with public access permission (https://fileshare.czbiohub.org/s/waXQzQ9PRZPwTdk) and can be accessed through GitHub (https://github.com/zjshi/gt-pro).

Code availability

The implementation and documentation of GT-Pro is available on the GitHub (https://github.com/zjshi/gt-pro). GT-Pro is written in C++ with python scripts, it is released as open-source software under the MIT license.

References

40. Almeida, A. et al. A unified sequence catalogue of over 280,000 genomes obtained from the human gut microbiome. Preprint at *bioRxiv* https://doi.org/10.1101/762682 (2019).
41. Nei, M. & Gojobori, T. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.* **3**, 418–426 (1986).
42. Chang, C. C. et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* **4**, 7 (2015).
43. Edgar, R. C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**, 2460–2461 (2010).
44. Liu, X. et al. A novel data structure to support ultra-fast taxonomic classification of metagenomic sequences with *k*-mer signatures. *Bioinformatics* **34**, 171–178 (2017).
45. Kokot, M., Długosz, M. & Deorowicz, S. KMC 3: counting and manipulating *k*-mer statistics. *Bioinformatics* **33**, 2759–2761 (2017).
46. Mende, D. R., Sunagawa, S., Zeller, G. & Bork, P. Accurate and universal delineation of prokaryotic species. *Nat. Methods* **10**, 881–884 (2013).
47. Hyatt, D. et al. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinf.* **11**, 119–119 (2010).
48. Kultima, J. R. et al. MOCAT: a metagenomics assembly and gene prediction toolkit. *PLoS ONE* **7**, e47656 (2012).
49. Sievers, F. et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* **7**, 539 (2011).
50. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
51. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
52. Segata, N. et al. Metagenomic microbial community profiling using unique clade-specific marker genes. *Nat. Methods* **9**, 811–814 (2012).
53. Gourlé, H., Karlsson-Lindsjö, O., Hayer, J. & Bongcam-Rudloff, E. Simulating Illumina metagenomic data with InSilicoSeq. *Bioinformatics* **35**, 521–522 (2018).
54. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinforma.* **30**, 1312–1313 (2014).
55. Olm, M. R., Brown, C. T., Brooks, B. & Banfield, J. F. dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. *ISME J.* **11**, 2864–2868 (2017).
56. Camacho, C. et al. BLAST+: architecture and applications. *BMC Bioinf.* **10**, 421 (2009).

Acknowledgements

This study was funded by NSF (grant #1563159), Chan Zuckerberg Biohub, Chan Zuckerberg Initiative, and Gladstone Institutes.

Author contributions

K.S.P. and S.N. conceived the project. K.S.P., S.N. and Z.J.S. designed experiments and drafted the manuscript. Z.J.S. conducted experiments, analyzed data, made figures and wrote software. B.D. wrote software and contributed to analysis of software performance. C.Z. contributed to analysis of structural variation imputation and tested software. K.S.P. supervised the project, provided computational resources and funding. K.S.P. and S.N. provided feedback. All authors read, edited and reviewed the paper.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41587-021-01102-3.

Correspondence and requests for materials should be addressed to Stephen Nayfach or Katherine S. Pollard.

Peer review information *Nature Biotechnology* thanks Yun William Yu, Falk Hildebrand and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

nature research

Double-blind peer review submissions: write DBPR and your manuscript number here

Corresponding author(s): Katherine S. Pollard

Last updated by author(s): 2021-08-09

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	We used public data as described in the methods section and supplementary tables. The Genome assemblies used in this study were downloaded conventionally. All metagenomic samples were downloaded from NCBI Sequence Read Archive (SRA). We notably used sratoolkit (version 2.10.0) to facilitate the retrieval of these data.
Data analysis	We used open source and custom code as described in the main text and methods section. Here is a list of software used in this study: MUMmer (version 4.0.0beta2), PLINK (version 1.9), KMC (version 3.1.1), Barrnap (version 0.9), Prodigal (version 2.6.3), fetchMG (version v1.0), Clustal Omega (version 1.2.4), FastTree2 (version 2.1.10), cophenetic.phylo from the ape package (version 5.3) in R, metaSNV (September 2019), BWA (version 0.7.17-r1188), SAMtools (version 1.9), MIDAS (September 2019), Bowtie2 (version v2.3.2), PySAM (version 0.12.0.1), StrainPhlAn (version 3), Kraken2 (version 2.0.9-beta), SRA toolkit(version 2.10.0), InSilicoSeq (version 1.4.2), checkM (version 1.0.11), dRep (version 1.4.3), BLAST (version 2.7.1), randomForest package (version 4.6-14) in R, umap package (version 0.2.3.1) in R, RAxML (version 8.2.12). Our GT-Pro code is available at https://github.com/zjshi/gt-pro.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All datasets are publicly available through the corresponding repositories. Genome assemblies for building GT-Pro used in this study were downloaded from the UHGG database and are available at MGnify (http://ftp.ebi.ac.uk/pub/databases/metagenomics/mgnify_genomes). The 1,171 C. difficile genomes are available at NCBI RefSeq (https://ftp.ncbi.nlm.nih.gov/genomes/refseq/bacteria/Clostridioides_difficile/), the accession numbers of 114 high-quality non-redundant C. difficile genomes are in Table S1S. All metagenomic samples are available at NCBI Sequence Read Archive (SRA; https://www.ncbi.nlm.nih.gov/sra) with accession numbers in supplementary tables: 2S,133 human microbiome samples (Table S8), Tanzania (Table S9), North America (Table S10), Madagascar (Table S11), and North American IBD cohort (Table S12) and global biogeography samples (Table S13). The GT-Pro SNP databases and genotype profiles of 2S,133 human microbiome samples generated in this study are available in a cloud server with public access permission (https://fileshare.czbiohub.org/s/waXQzQ9PRZPwTdk) and can be accessed through GitHub (https://github.com/zjshi/gt-pro).

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	We sampled a 909 species and 112,904 genomes from UHGG database. The sample size was determined by the availability of high-quality genomes in UHGG database.
Data exclusions	Genomes with lower quality than high-quality (completeness >= 90% and contamination rate <= 5%) genomes were excluded. Species with fewer than 10 high-quality (completeness >= 90% and contamination rate <= 5%) whole genome sequences. Species with more than 1000 genomes were further downsized to only the top 1000 genomes in terms of quality metrics and genome length.
Replication	Not applicable - no data was generated in the study. We used all biological replicates of genomes and metagenomes in public data.
Randomization	Not applicable - no data was generated in the study. We used a random number generator in the simulations.
Blinding	Not applicable - no data was generated in the study, and all data is publicly available.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging