nature portfolio

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

Supplementary information

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

In the format provided by the authors and unedited

Supplementary figures for "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,*]

[1]Chan Zuckerberg Biohub, Data Science, San Francisco, CA, [2]Gladstone Institutes, San Francisco, CA, [3]Chan Zuckerberg Initiative, CA, [4]Department of Energy, Joint Genome Institute, Walnut Creek, CA, [5]Lawrence Berkeley National Laboratory, Environmental Genomics and Systems Biology Division, Berkeley, CA, [6]University of California San Francisco

*e-mail: snayfach@lbl.gov; katherine.pollard@gladstone.ucsf.edu

Supplementary figures

Figure S1

Implementation of the GT-Pro *in sillico* metagenotyping framework for the human gut microbiome (layout mirroring Figure 1). We identified 909 gut species with at least 10 high-quality genomes in the UHGG. More than 104 million common SNPs were called through whole genome alignment of conspecific genomes. Up to 124 (31 X 4) candidate k-mers were extracted per bi-allelic SNP site, a total of 1.2 trillion candidate SNP-covering k-mers (sck-mers) were extracted, and among them 5.7 billion sck-mers were determined to be species-specific. The resulting GT-Pro database directly covers 52.8 million SNPs from 881 species. In parallel, a total of 6.8 million LD blocks were detected by clustering SNPs based on pairwise linkage disequilibrium (LD). Within LD blocks, about 15 million SNPs were assigned as tag SNPs, including standalone SNPs (not in LD blocks with any other SNP). Tag SNPs must be covered by at least one species-specific sck-mer for each allele. An additional 15.7 million "indirectly covered" SNPs without sck-mers are in strong LD ($r2 > 0.81$) with a tag SNP. Database characteristics are compared to those of Schlossnig et al[31]. and the 1000 genome project catalog of human genetic diversity.

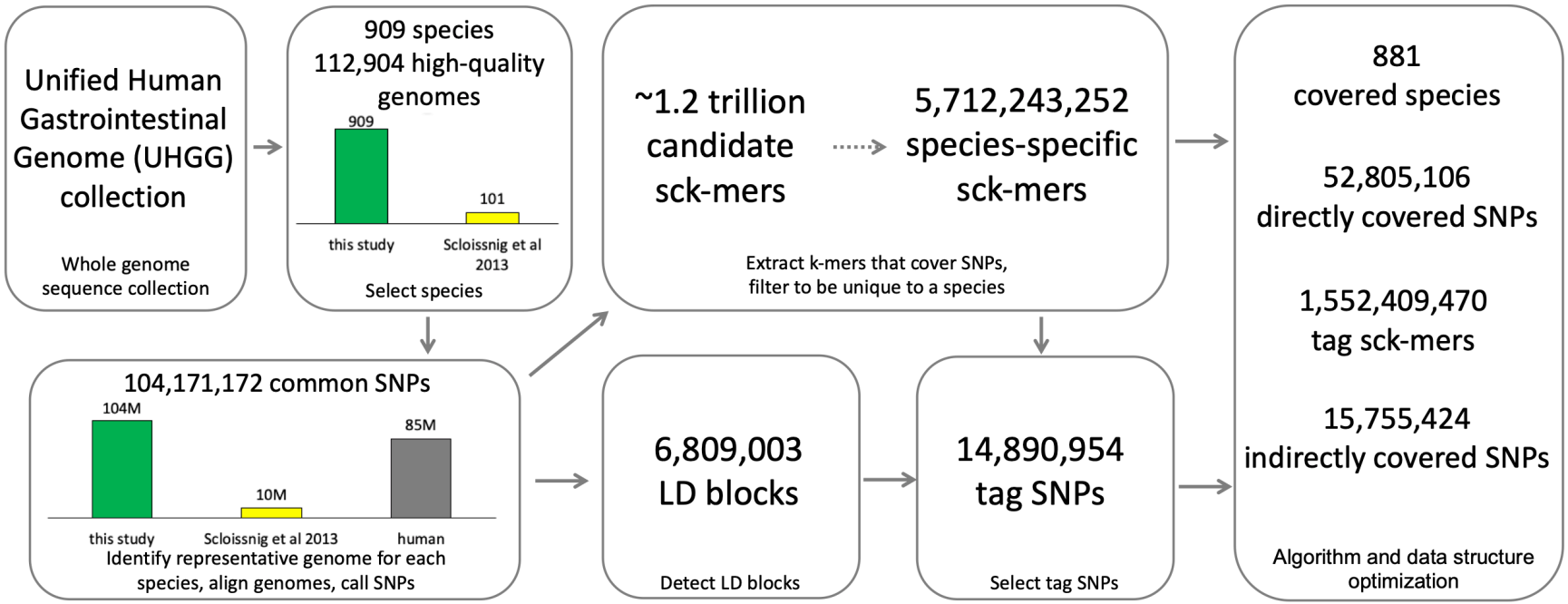


Figure S2

Distribution of genomes used to build the GT-Pro human gut database according to their quality metrics, including (a) completeness and (b) contamination rate. The black dashed lines represent the mean from the representative genomes.

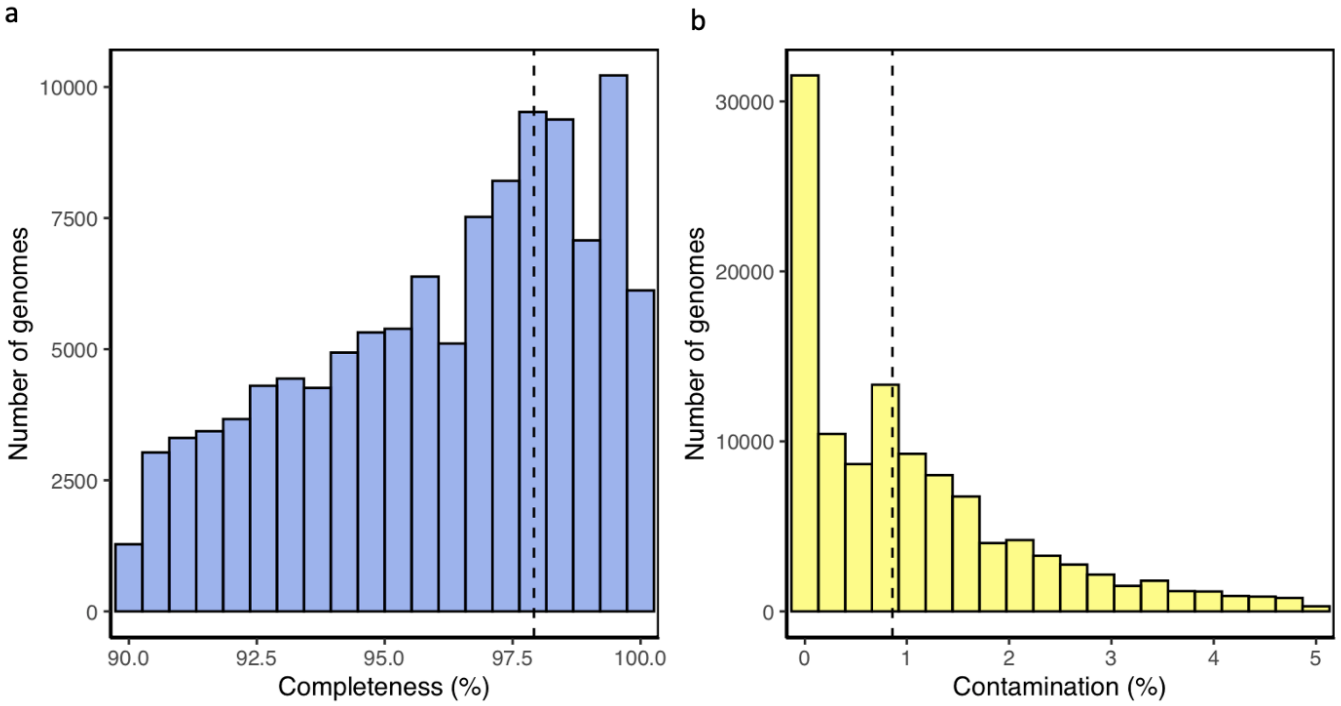
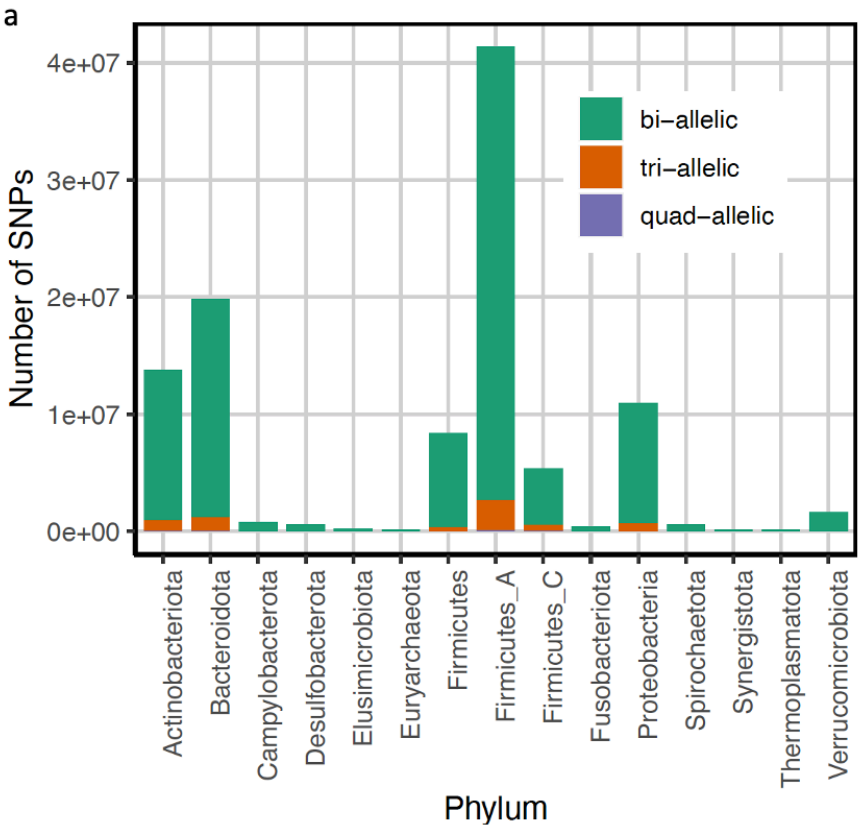


Figure S3

(a) Counts of bi-, tri- and quad-allelic common SNPs discovered in 909 genomes, summarized across gut bacterial phyla. (b) Number of SNPs per bacterial phylum colored by being in GT-Pro (green), in an LD block with a SNP in GT-Pro (orange), or not detectable by GT-Pro (purple).



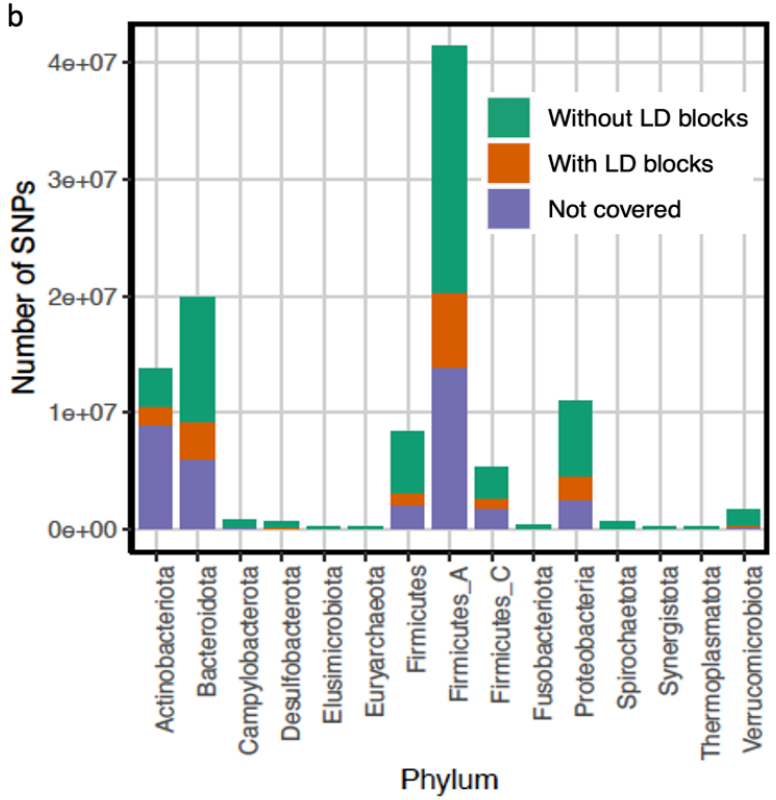


Figure S4
Characteristics of common SNPs (from top to bottom) across species ordered by phylum and number of SNPs: number of discovered alleles, location of bi-allelic SNPs in coding or non-coding sequences, proportion of coding mutatinos that are synonymous versus non-synonymous, and types of non-synonymous mutation (non-sense, stop codon disruptive or non-stop codon).

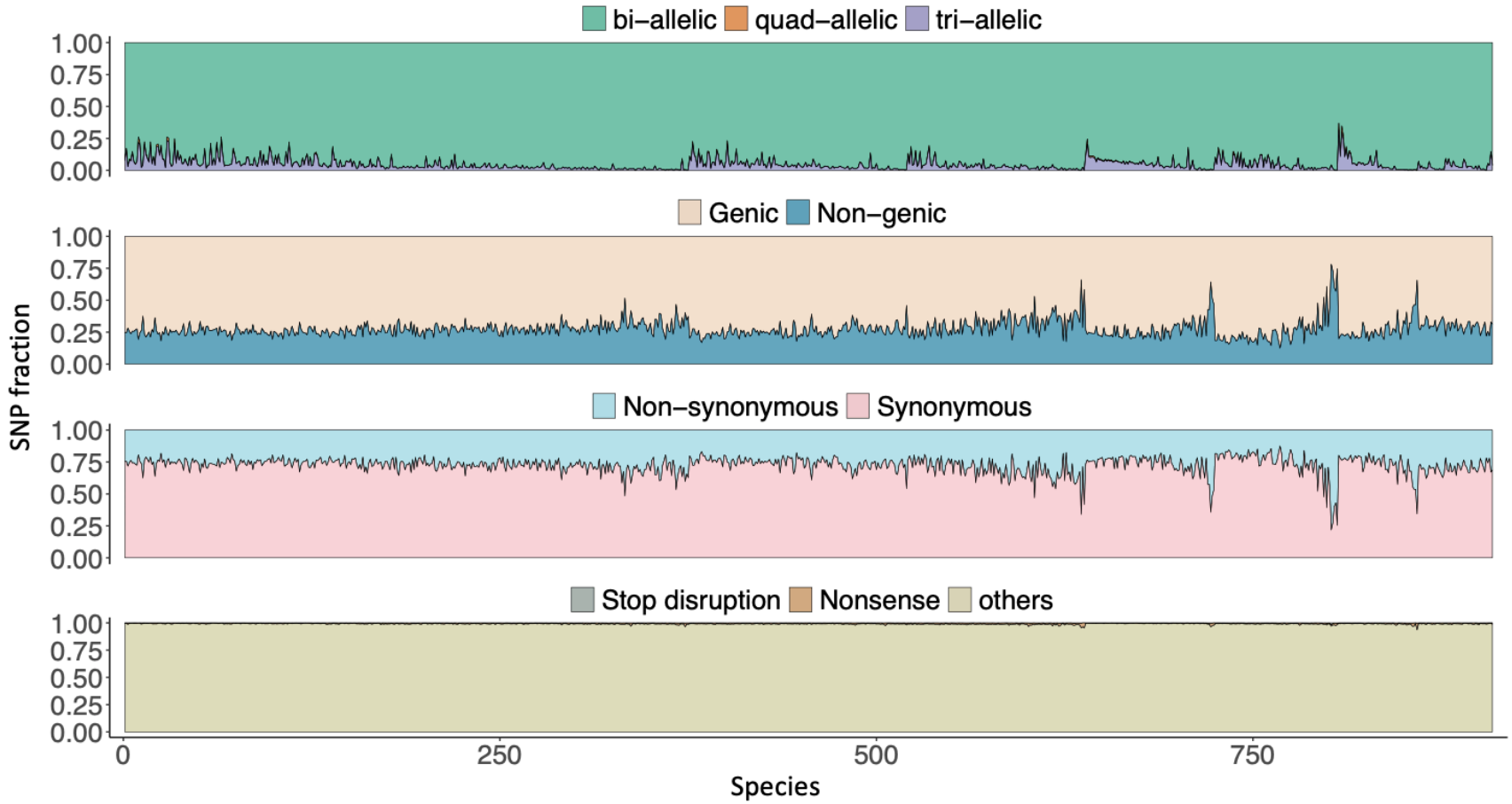


Figure S5
(a and b) Correlation between the number of genomes and (a) the number of discovered SNPs or (b) the number of SNPs per Kb. The black dashed lines in both (a) and (b) are linear regression lines for each phylum. (c) Comparison of average LD block size (e) across gut bacterial phyla.

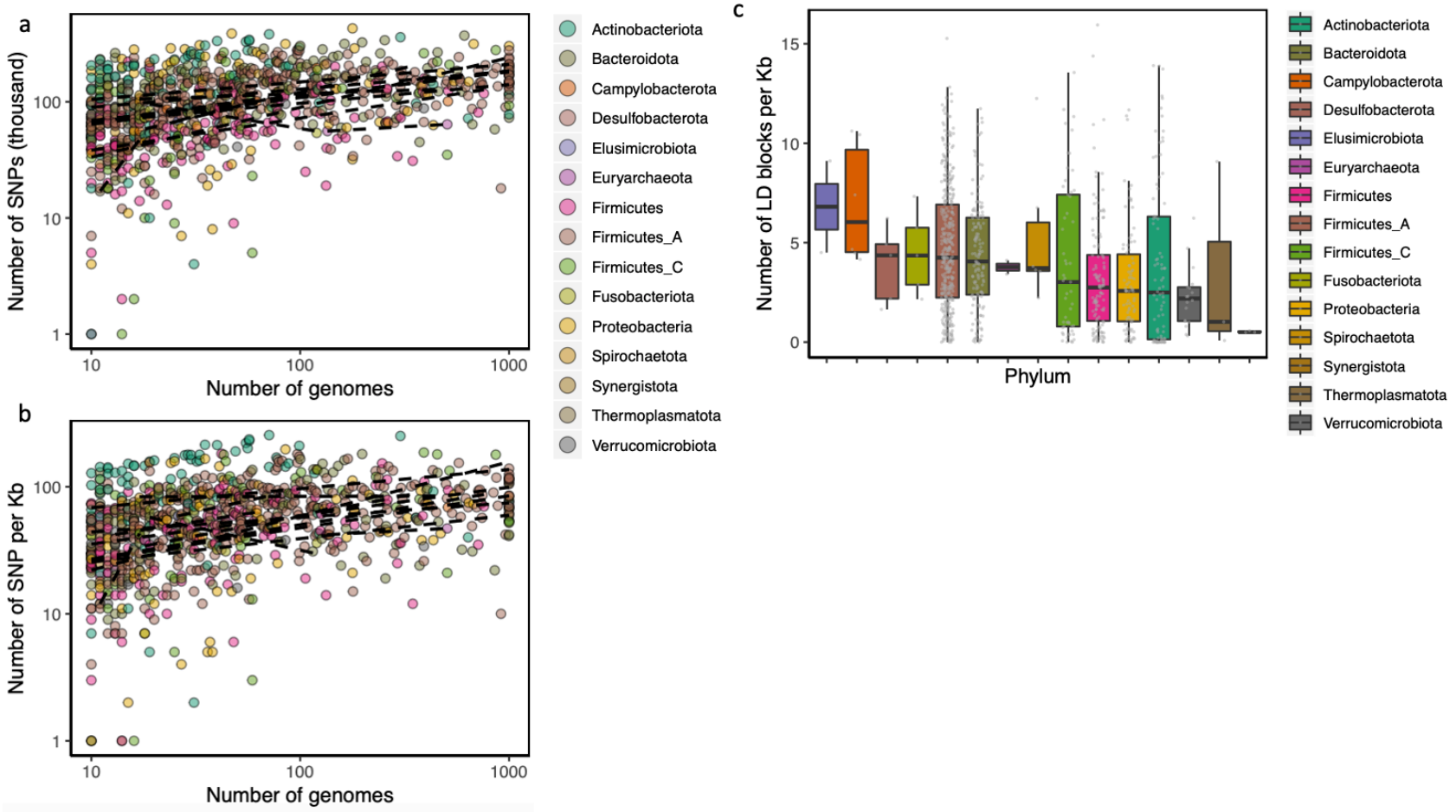


Figure S6
(a) Rarefaction curves of common SNP discovery for the species with at least 50 genomes. Each curve represents a species and is made by calculating the number of SNPs that can be discovered using down-sampled subsets of genomes per species. The subsets start with 10 genomes and increase by 10 genomes up to the true number. This down-sampling was repeated 10 times and the mean number of SNPs discovered at each subset size is plotted. (b) Histogram showing the percent increase in common SNPs discovered when adding the last 10 genomes for each species in (a). Most species have a low rate of increase, indicating that SNP discovery has leveled off. But a few species have 8-10% increases, suggesting that more common SNPs remain to be discovered.

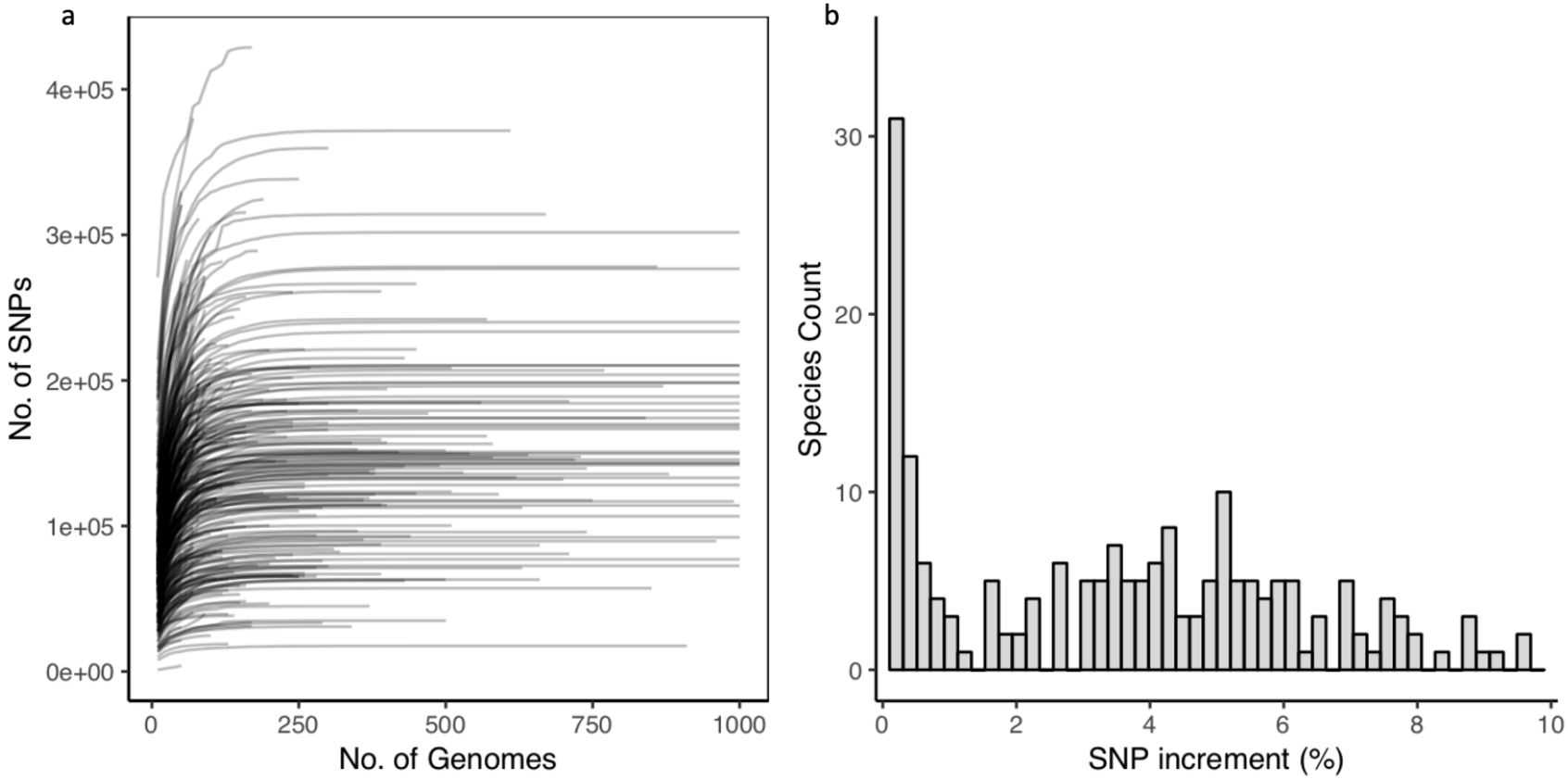


Figure S7
Correlation between the number of SNPs per Kb and the nucleotide diversity. Each dot in the plot represents a species and dot color reflects the phylum of the species. The black dashed line is the linear regression line.

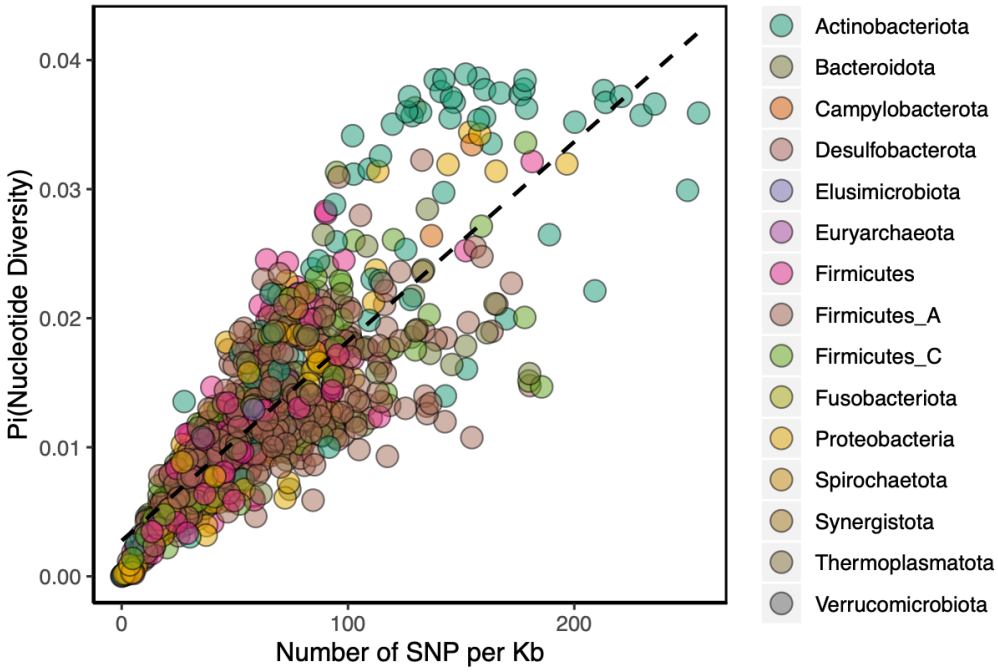


Figure S8

Intra-specific pairwise SNP density in the phylum of *Actinobacteriota*. Each bar in the plot represents a species (denoted by its ID number) and the height of the bar indicates the median of pairwise SNP density in a species.

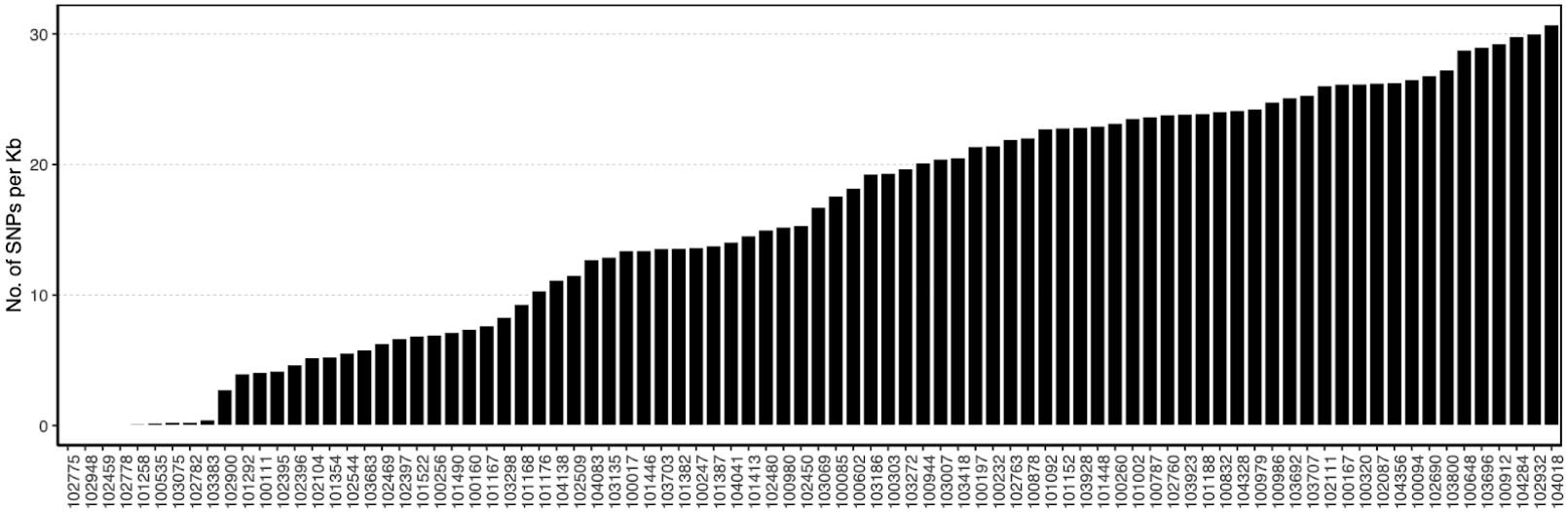


Figure S9
Distribution of pN/pS per species across gut bacterial phyla. The dots above the black horizontal dashed line represent the species with both high pN/pS (> 0.5) and low pS (< 0.007).

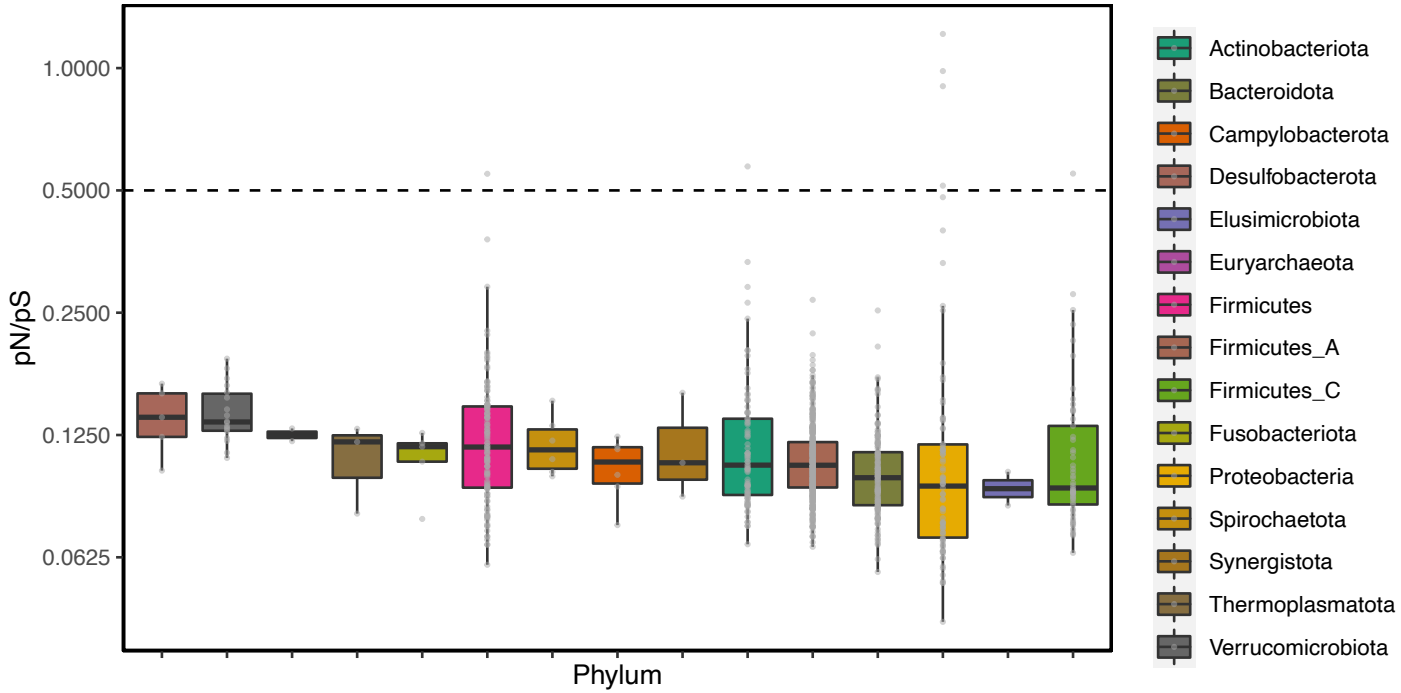


Figure S10
(a) Summary of average number of species-specific sck-mers per SNP for all species. Horizontal dashed lines and the bar on the right-side group species by average number, with darker red indicating species with more sck-mers per SNP. (b) Correlation between average number of species-specific sck-mers per SNP and the number of genomes for each species. Each dot in the plot represents a species. The black dashed line is the linear regression line.

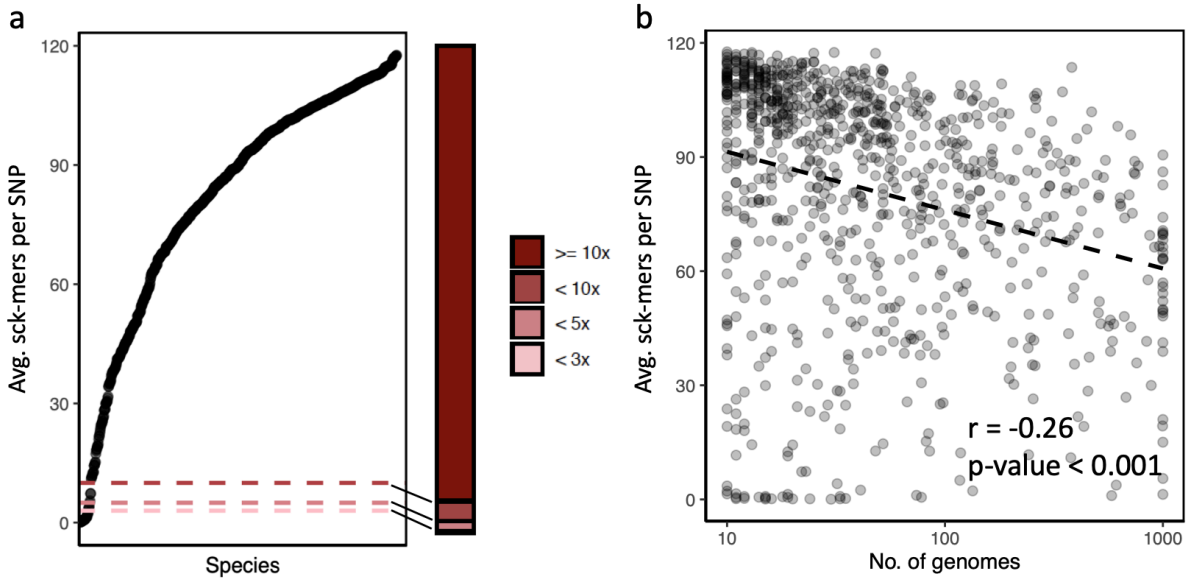


Figure S11
Resolution of within-species genetic diversity captured by different subsets of SNPs. The genetic distances (phylogenetic tree branch length by FastTree 2.0) between all pairs of conspecific genomes were computed using five sets of SNPs: entire catalogue of common SNPs regardless of having sck-mers (all), all SNPs in the GT-Pro database (full), GT-Pro tag SNPs (tag), only SNPs in 40 universal marker genes (marker), and only SNPs in the 16S rRNA gene (16S). (a) Distances using each subset of common SNPs (rows) compared to pairwise ANI (includes rare SNPs) for eight species (columns) with the most genomes. Species (from left to right): *Agathobacter rectalis* (genus *Agathobacter*) (species id: 102492), *Bifidobacterium infantis* (genus: *Bifidobacterium*) (species id: 101292), an unassigned species from *Ruminococcus_E genus* (species id: 100258), *Klebsiella pneumoniae* (genus: *Klebsiella*) (species id: 102538), *Vibrio cholerae* (genus: *Vibrio*) (species id: 102311), *Salmonella enterica* (genus *Salmonella*) (species id: 102366), *Campylobacter jejuni* (genus: *Campylobacter*) (species id: 102422), an unassigned species from *Escherichia* genus (species id: 102506). (b) Heatmap of Spearman's rank correlations between pairwise ANIs and the five common SNP-based distances (rows) for all 909 gut bacterial species (columns, sorted from low to high correlation). (c) Boxplot summarizing the distributions of correlations in each row of (b). The correlation of 16S and whole genome ANI increases (median=0.28, IQR=0.44) when only cultured isolates are used.

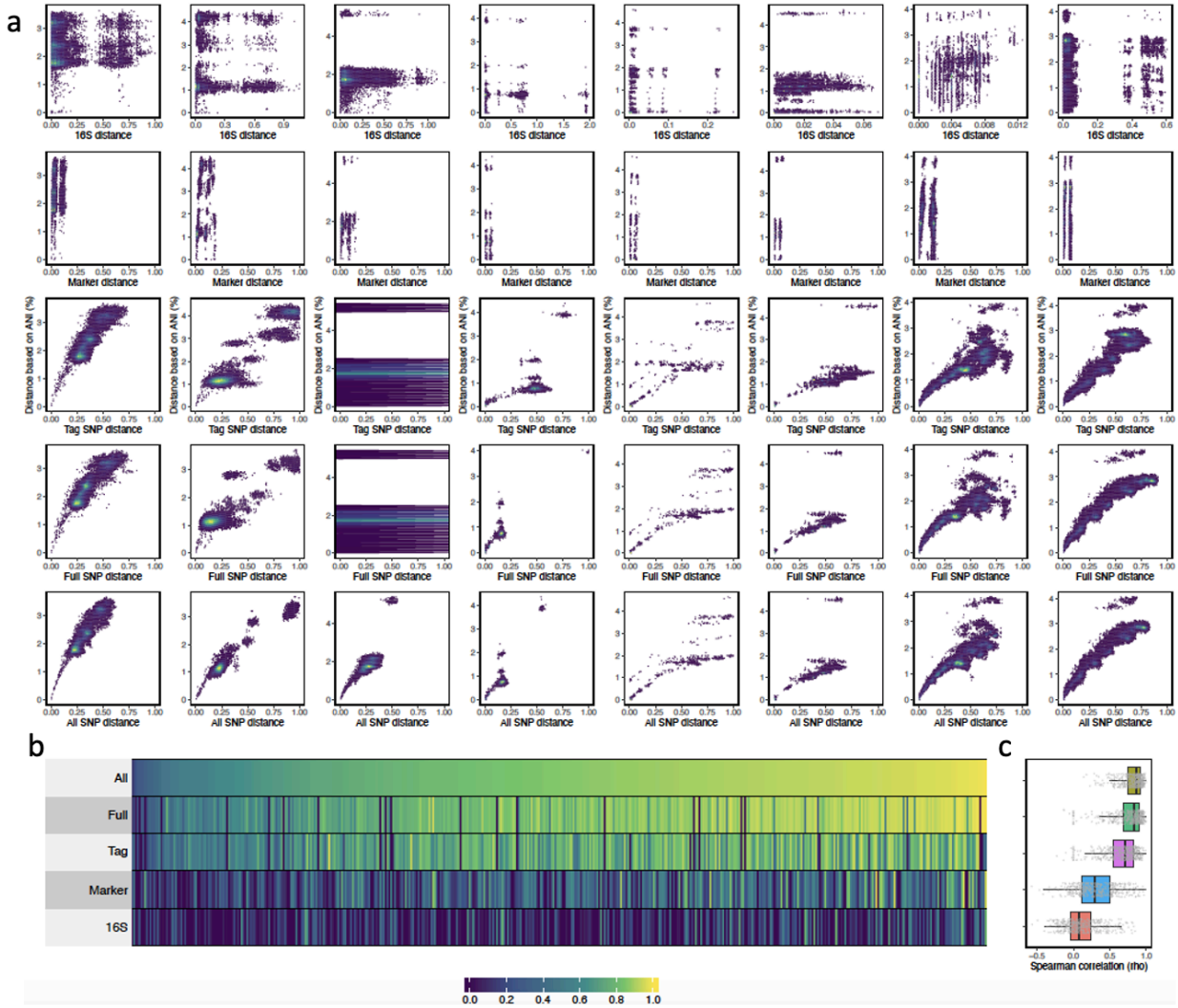


Figure S12

Examples for overlap encoding in the sc-span table. (a) An example of a sc-span entry which only has a single SNP (red). Two underscored sequences are the full-length (61bp) sc-spans for both major and minor allele. (b) An example of a sc-span entry which has two SNPs (red, blue). A third full-length sc-span (underscored and labeled as sc-span') was introduced to resolve the additional SNP without duplicating data uneccessarily by creating a span for each SNP.

a

SNP center	SNP offset
..............................GGTCTAGGCGCAATGTAACGCTTTTATCGCT	0
..........................TCTTGGTCTAGGCGCAATGTAACGCTTTTAT....	4
....................GCTTAGTCTTGGTCTAGGCGCAATGTAACGC..........	10
..............TTAGCGCTTAGTCTTGGTCTAGGCGCAATGT..............	15
............GACTTAGCGCTTAGTCTTGGTCTAGGCGCAA..................	18
.........CTAGACTTAGCGCTTAGTCTTGGTCTAGGCG.....................	21
CAACTATTCCTAGACTTAGCGCTTAGTCTTGGTCTAGGCGCAATGTAACGCTTTTATCGCT	Major sc-span
CAACTATTCCTAGACTTAGCGCTTAGTCTTCGTCTAGGCGCAATGTAACGCTTTTATCGCT	Minor sc-span
....................GCTTAGTCTTCGTCTAGGCGCAATGTAACGC..........	10
..............CTTAGCGCTTAGTCTTCGTCTAGGCGCAATG..............	16
.........CTAGACTTAGCGCTTAGTCTTCGTCTAGGCG.....................	21
.....ATTCCTAGACTTAGCGCTTAGTCTTCGTCTA..........................	25
....TATTCCTAGACTTAGCGCTTAGTCTTCGTCT...........................	26

b

SNP center	SNP offset
......................•.....TCTTGGTCTAGGCGCAATGTAACGCTTTTAT....	4
....................GCTTAGTCTTGGTCTAGGCGCAATGTAACGC..........	10
..............TTAGCGCTTAGTCTTGGTCTAGGCGCAATGT..............	15
............GACTTAGCGCTTAGTCTTGGTCTAGGCGCAA..................	18
.........CTAGACTTAGCGCTTAGTCTTGGTCTAGGCG.....................	21
CAACTATTCCTAGACTTAGCGCTTAGTCTTGGTCTAGGCGCAATGTAACGCTTTTATCGCT	Major sc-span
CAACTATTCCTAGACTTAGCGCTTAGTCTTCGTCTAGGCGCAATGTAACGCTTTTATCGCT	Minor sc-span
....................GCTTAGTCTTCGTCTAGGCGCAATGTAACGC..........	10
..............CTTAGCGCTTAGTCTTCGTCTAGGCGCAATG..............	16
CAACTATTCCTAGACTTAGCTCTTAGTCTTCGTCTAGGCGCAATGTAACGCTTTTATCGCT	Minor sc-span'
.........CTAGACTTAGCTCTTAGTCTTCGTCTAGGCG.....................	21
.....ATTCCTAGACTTAGCTCTTAGTCTTCGTCTA..........................	25
....TATTCCTAGACTTAGCTCTTAGTCTTCGTCT...........................	26

Figure S13
Summary of number of SNPs (a) and sck-mers (b) in full and tag version of GT-Pro database. Each observation is one species.

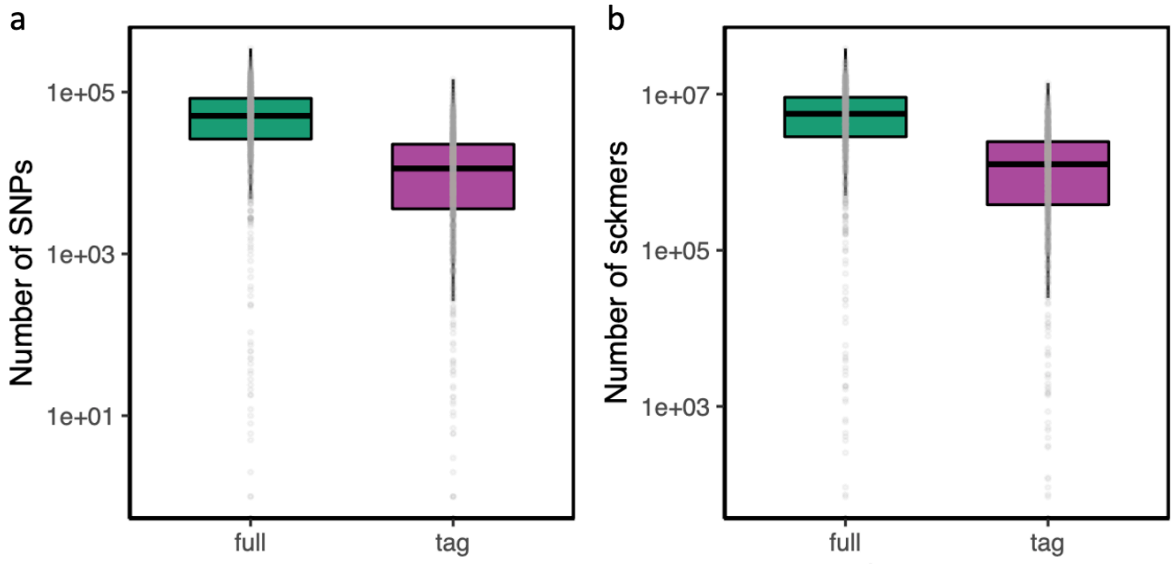


Figure S14

Examples of LD distance decay as genomic distance between SNPs increases: (a) Streptococcus salivarius (genus: Streptococcus) (species id: 100113), (b) Bacteroides thetaiotaomicron (genus: Bacteroides) (species 100196), (c) an unclassified species from Escherichia genus (species id: 102506) , (d) Roseburia inulinivorans (genus: Roseburia) (species id: 100271), (e) Veillonella parvula (genus: Veillonella) (species id: 101349), (f) Vibrio parahaemolyticus (genus: Vibrio) (species id: 102331), (g) Acinetobacter baumannii (genus: Acinetobacter) (species id: 102344), (h) GCA_001916965.1 (genus: COE1) (species id: 104431).

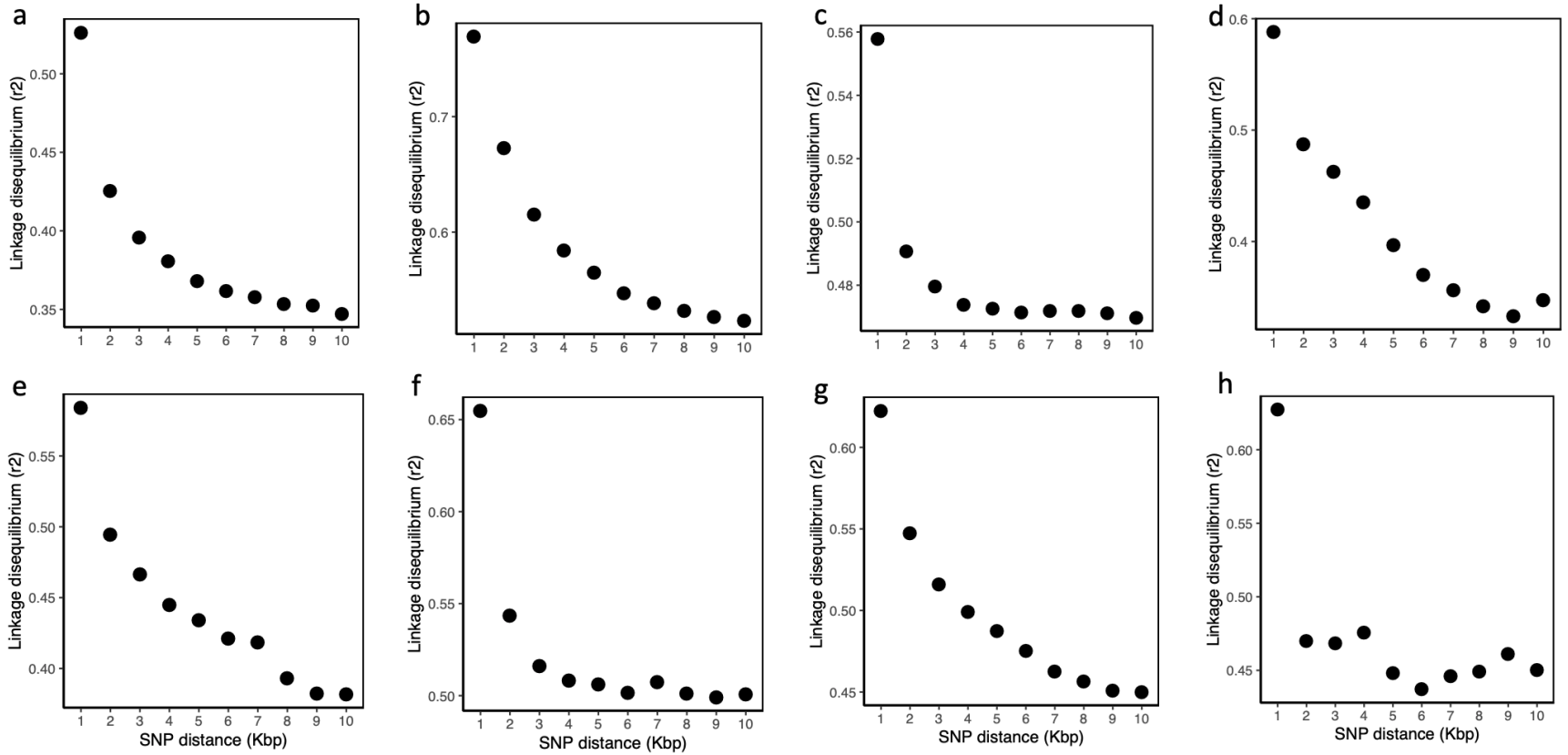


Figure S15
Number of LD blocks identified with increasing r2 cutoff. Each observation represents a species.

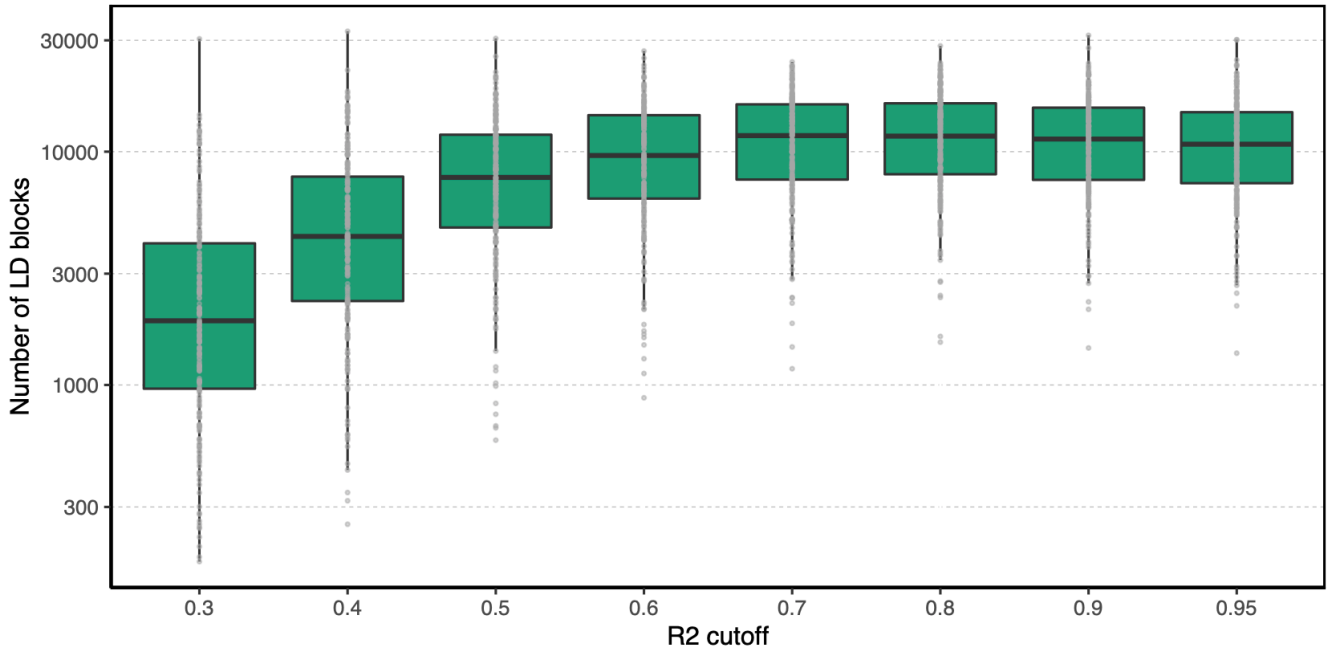


Figure S16
Distribution of LD blocks per Kb across species, including all singleton SNPs as LD blocks. Red vertical line: one LD block per Kb.

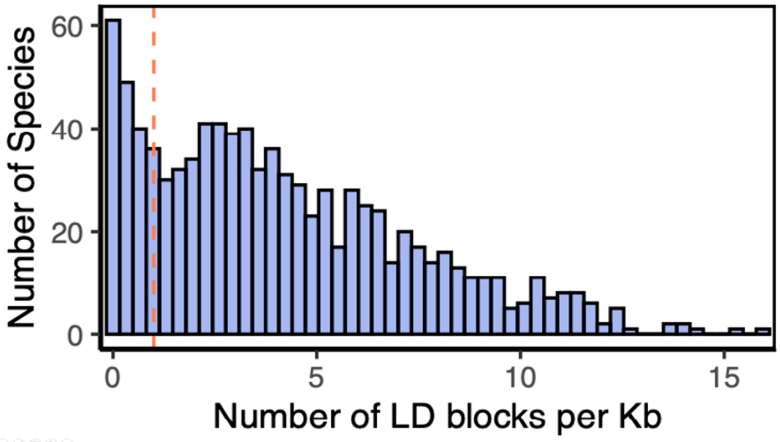


Figure S17

Development and optimization of GT-Pro species-specific sck-mer data structures and algorithms. (a) Workflow for metagenotyping. In the GT-Pro database's sck-mer table, all species-specific sck-mers have been binary encoded, pooled, and sorted in Colex order. The bits of sck-mers were split into M-bit prefixes for a quick filter (M-filter; blue) and L-bit suffixes to build an index (L-index; green) whose entries point to consecutive records in the sck-mer table (violet). Efficient exact-matching of k-mers from metagenomes starts with extracting all k-mers present in each sequencing read (input). These are passed through the M-filter. Only k-mers passing the filter can be exact matches to sck-mers. To determine if they are, the range of entries in the L-index corresponding to each L-bit hit is queried with an exact-match algorithm. Dashed black arrows indicate processes occurring during database development, and solid gray arrows indicate the metagenotyping workflow. (b) Schematic showing details of the sck-mer table data structure. (c) Details of the SNP-centered spans (sc-span) table.

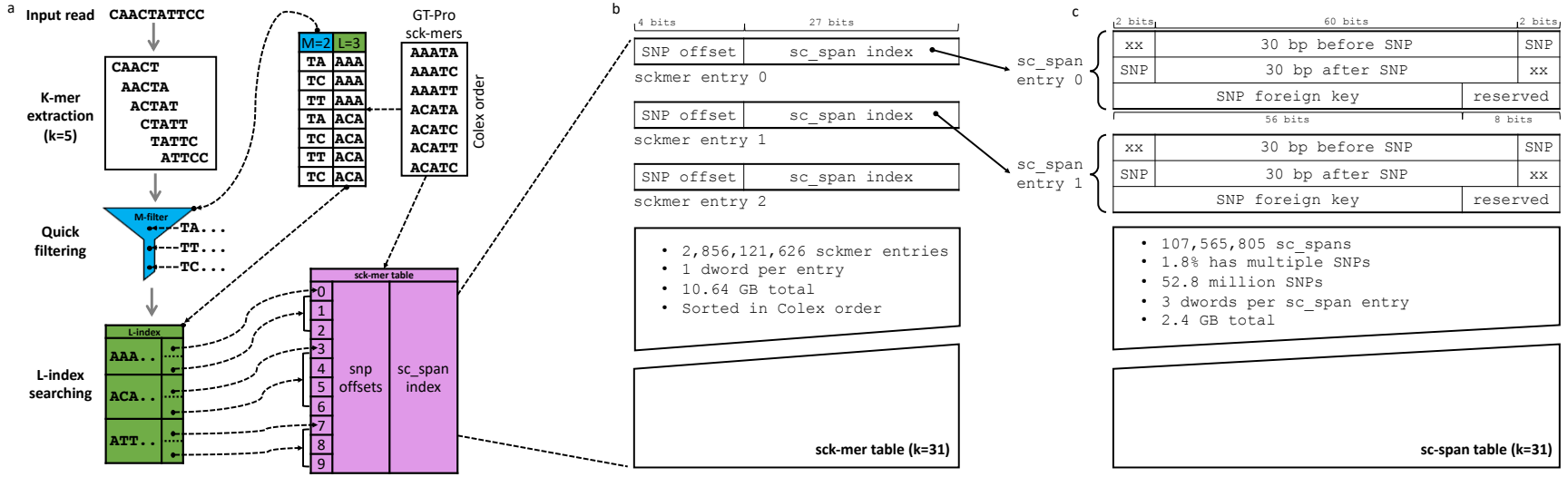


Figure S18

Accuracy comparison of GT-Pro genotypes from simulated forward and reverse reads. (a) False discovery rate of both reference and non-reference genotypes. (b) Sensitivity across coverage levels from the simulations in (a). (c) False discovery rate of non-reference genotypes. (d) Sensitivity across coverage levels from the simulations in (c).

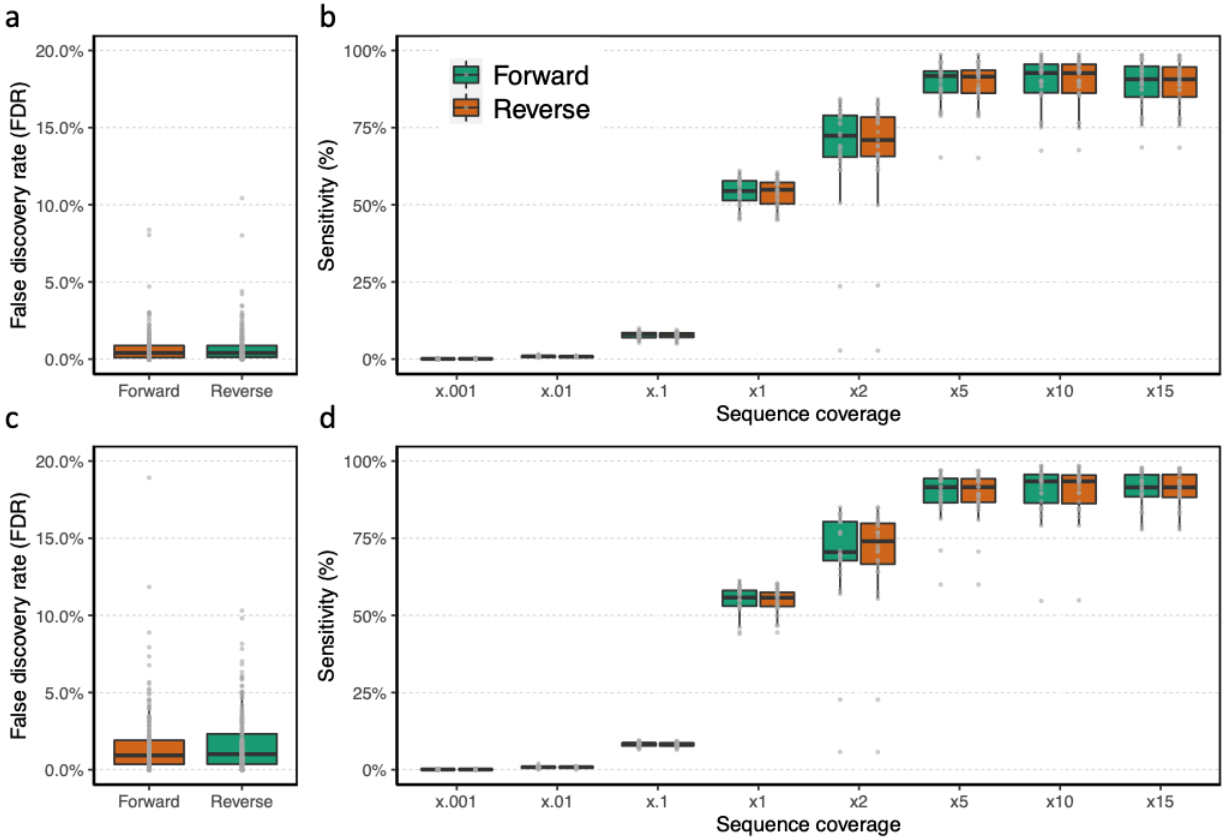


Figure S19

(a-b) Accuracy comparisons similar to **Figure 4**, except metagenotypes only include non-reference alleles compared to the representative genome for each species. (a) False discovery rate of non-reference metagenotypes at different sequencing coverages ranging from 0.001x to 15x. Each observation is the result from a metagenome containing reads from one isolate at one coverage value. False discoveries are genotype calls that do not match the genome from which reads were simulated. (b) Sensitivity across coverage levels from the simulations in (a). Sensitivity is the proportion of non-reference genotypes present in the isolate genome that are correctly detected by metagenotyping. (c-d) Similar to two-isolate simulations in **Figure 4**, except only using heterozygous sites where the two isolates have different alleles. (c) False discovery rate at heterozygous sites in metagenomes containing reads from two isolates of each species. Different ratios of sequencing coverage between pairs of isolates were simulated by fixing a more abundant isolate at 15x coverage in all simulations and varying the other isolate's coverage from 0.001x to 15x (coverage ratio = 0.001:15 to 15:15). (d) Sensitivity at heterozygous sites in metagenomes from (c). Sensitivity is the proportion of heterozygous present in the genomes from which reads were simulated that are correctly called in the metagenotypes.

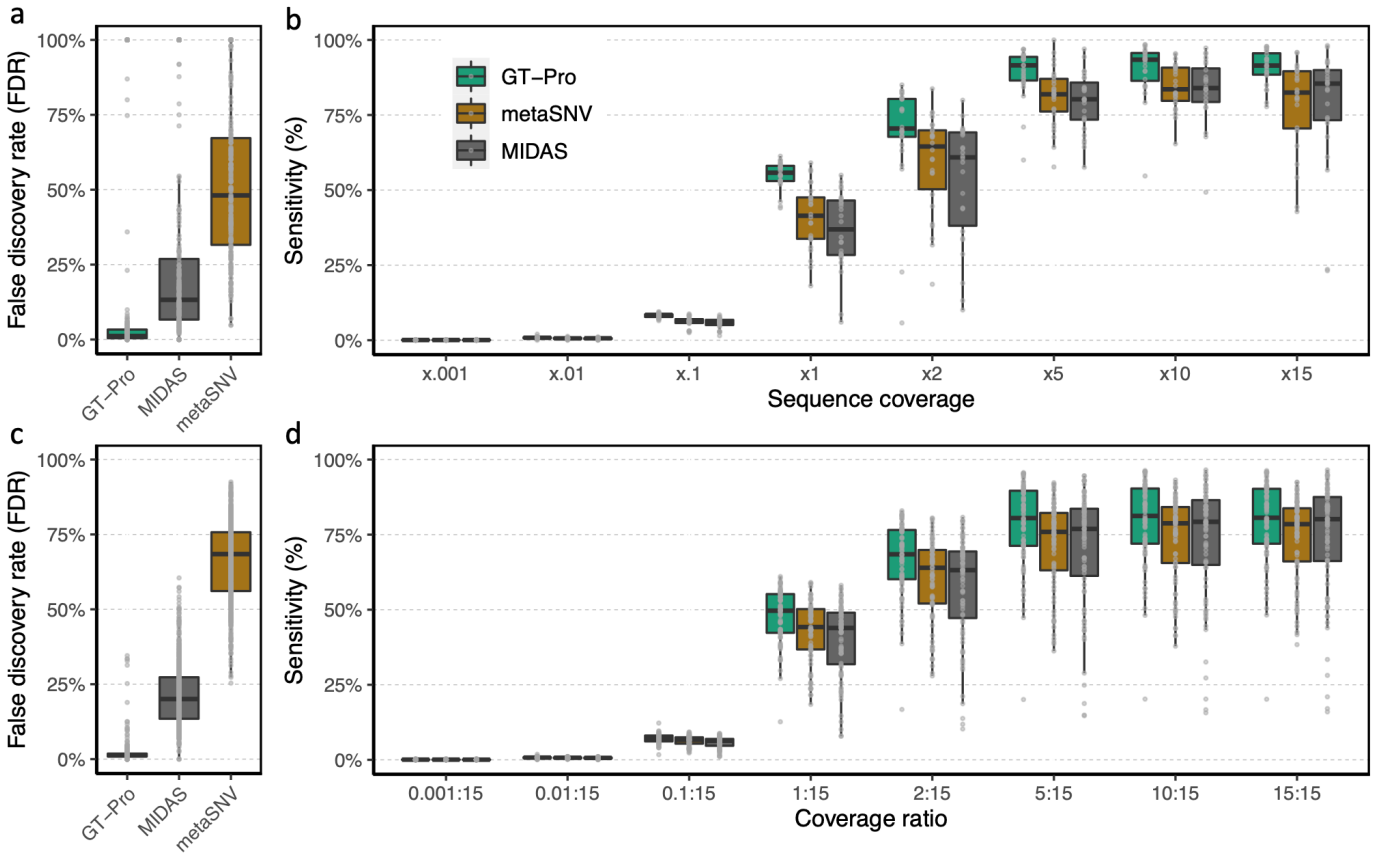


Figure S20
False discovery rate comparison of GT-Pro, MIDAS, and metaSNV at sites in GT-Pro database. False discovery rate of both reference and non-reference genotypes (a) and non-reference genotypes only (b) at a combination of sequencing coverages ranging from 0.001x to 15x.

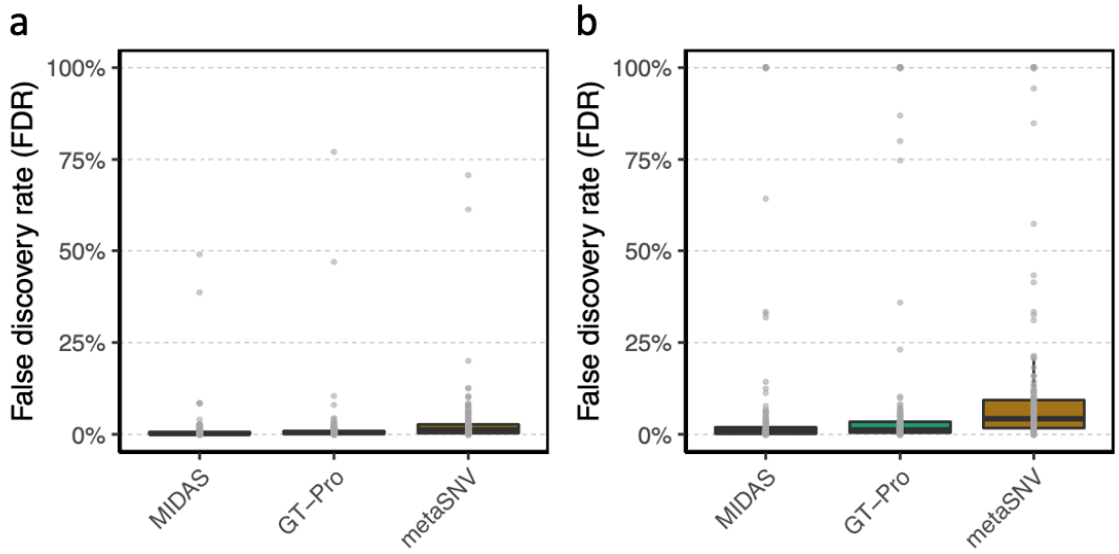


Figure S21
False discovery rate comparison of GT-Pro, MIDAS, and metaSNV at sites across whole genome. 5x coverage cutoff was applied to MIDAS and metaSNV. False discovery rate of both reference and non-reference genotypes (a) and non-reference genotypes only (b) at a combination of sequencing coverages ranging from 0.001x to 15x.

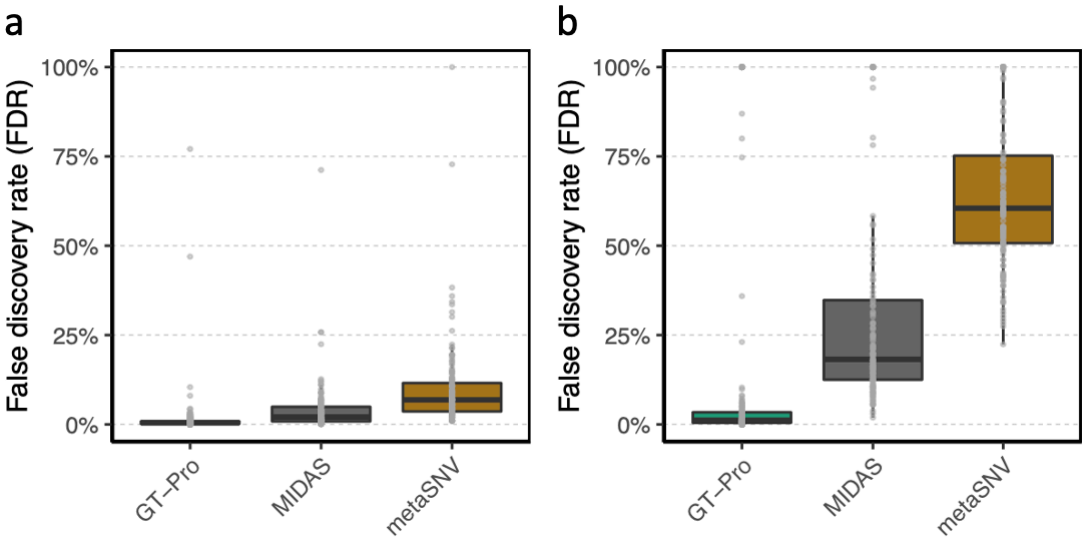


Figure S22
Comparison of median number of genotyped sites across detected species at a combination of sequencing coverages ranging from 0.001x to 15x. A 1x coverage cutoff was applied to GT-Pro. 1x and 5x coverage cutoffs were applied to both MIDAS and metaSNV. (a) Both reference and non-reference genotypes and (b) non-reference genotypes only.

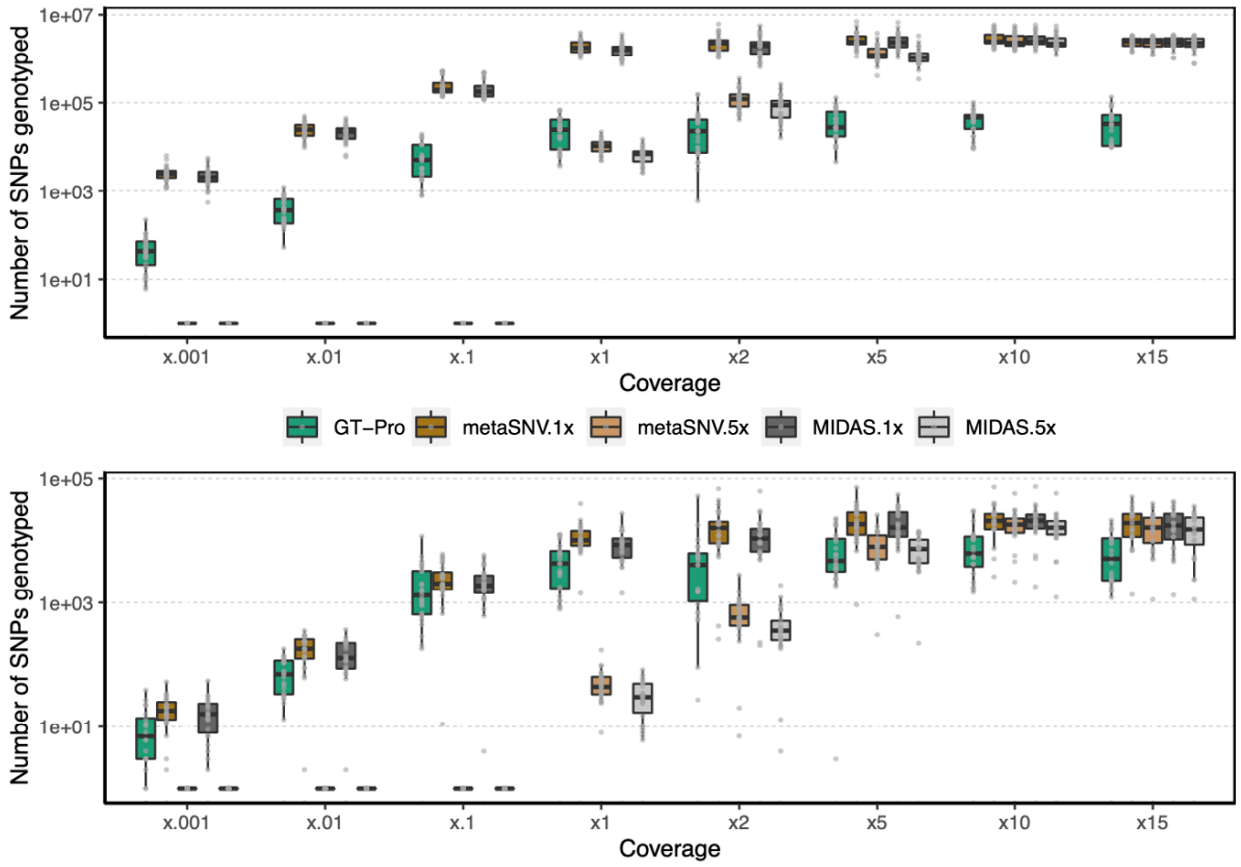


Figure S23
False discovery rate comparison of GT-Pro, 5-fold cross-validation GT-Pro, MIDAS, and metaSNV. False discovery rate of both reference and non-reference genotypes (a) and non-reference genotypes only (b) at a combination of sequencing coverages ranging from 0.001x to 15x.

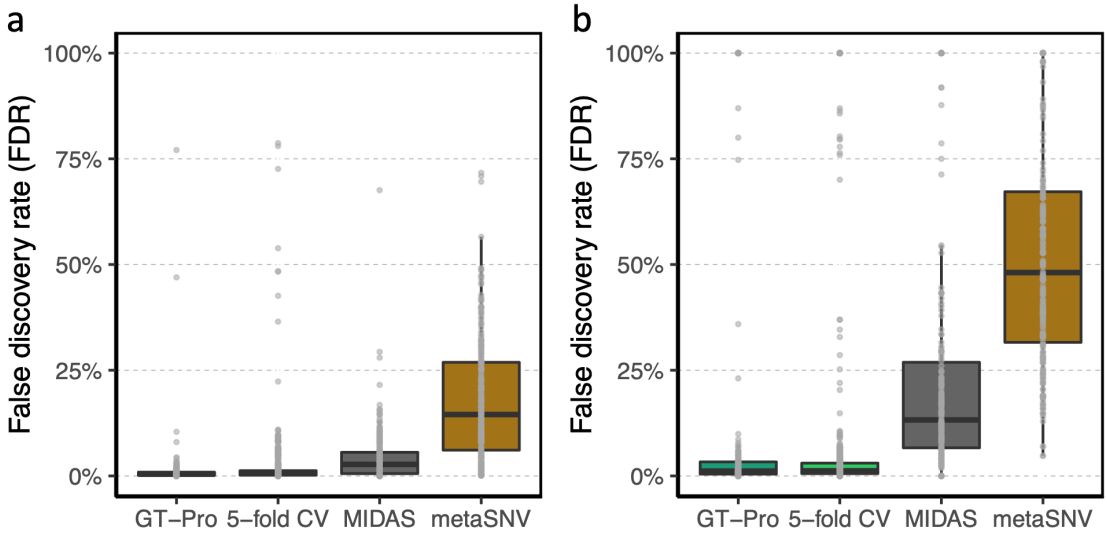


Figure S24
Distribution across species of odds ratios quantifying agreement between tag SNPs being genotyped and non-tag SNPs in the same LD block being genotyped (using the full version of the GT-Pro database). High odds ratios indicate species where non-tag SNPs are rarely absent if the tag SNP for their LD block is present. Odds ratios are based on genotyped sites in simulations in Figure 5a.

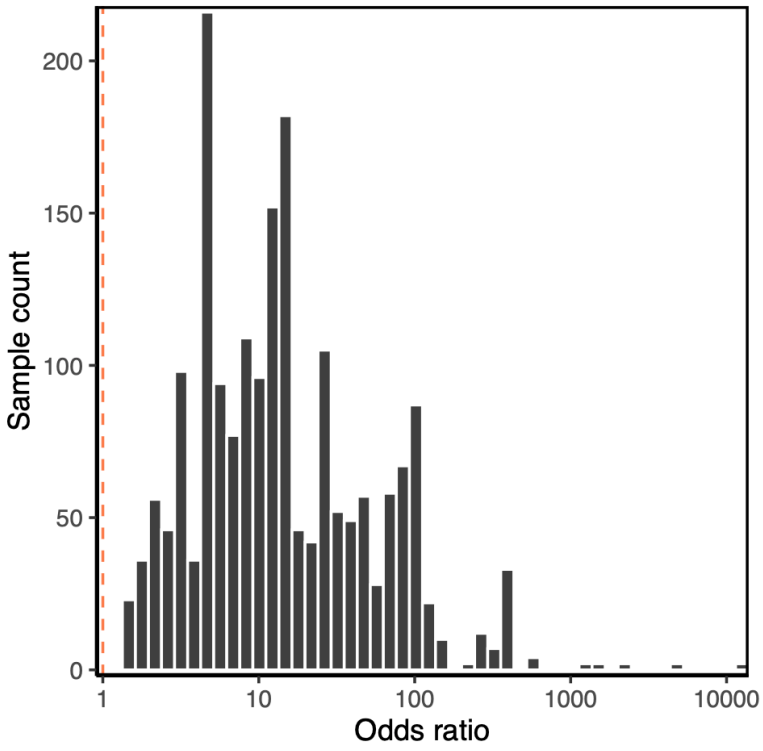


Figure S25

Misassignment rate comparison of GT-Pro and alignment methods. Misassignment rate of both reference and non-reference genotypes (a) and non-reference genotypes only (b) at a sequencing coverage of 15x. The genomes for simulation are the same as Fig 4a and b. No coverage cutoff was applied to MIDAS, metaSNV or GT-Pro. The rate of misassignment per species is the ratio between misassigned and total assigned genotypes and a genotype is determined to be misassigned to a species if it has one or more reads simulated from any other species (see Methods).

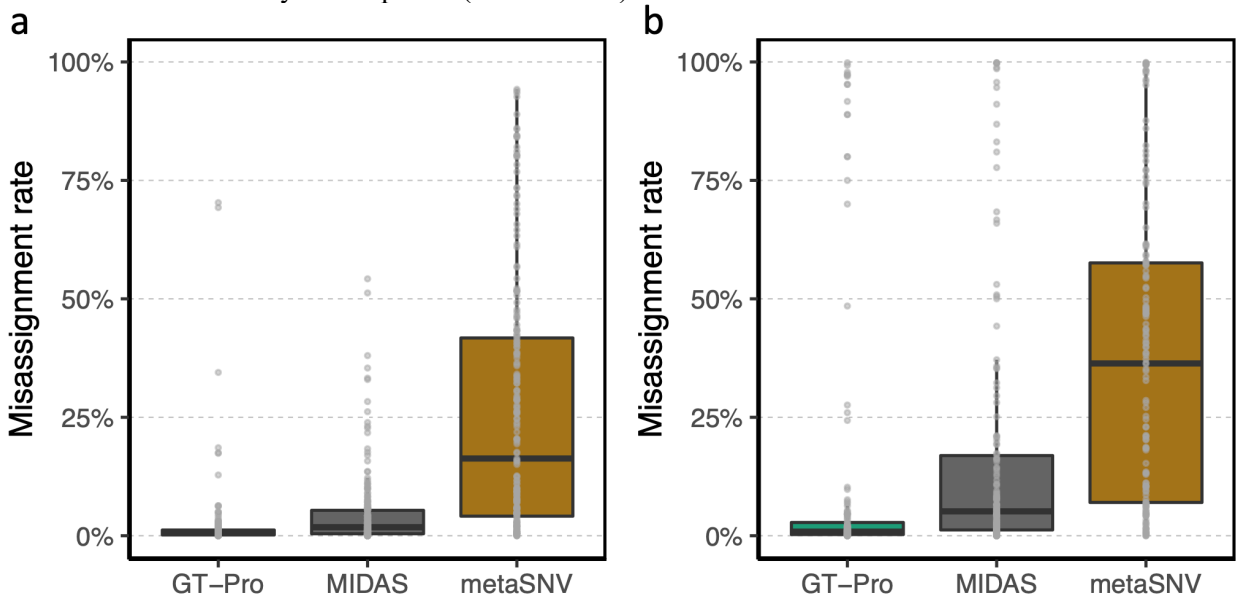


Figure S26
False discovery rate comparison of GT-Pro, MIDAS and metaSNV at both homozygous and heterozygous sites (a) and heterozygous sites only (b) in metagenomes containing reads from two isolates of each species. Only sites in GT-Pro database were included. A combination of sequencing coverage ratio between two isolates was simulated by fixing a more abundant isolate at 15x coverage in all simulations, and varying the other isolate's coverage from 0.001x to 15x (coverage ratio = 0.001:15 to 15:15).

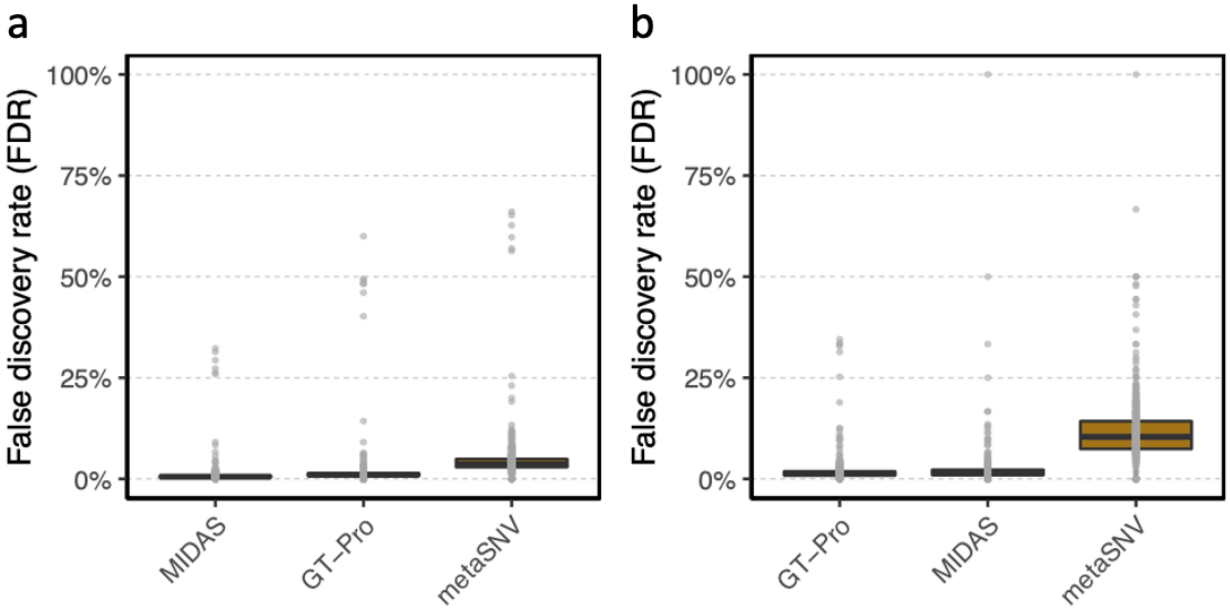


Figure S27
False discovery rate comparison of GT-Pro, MIDAS and metaSNV at both homozygous and heterozygous sites (a) and heterozygous sites only (b) in metagenomes containing reads from two isolates of each species. 5x coverage cutoff was applied to MIDAS and metaSNV. A combination of sequencing coverage ratio between two isolates was simulated by fixing a more abundant isolate at 15x coverage in all simulations, and varying the other isolate's coverage from 0.001x to 15x (coverage ratio = 0.001:15 to 15:15).

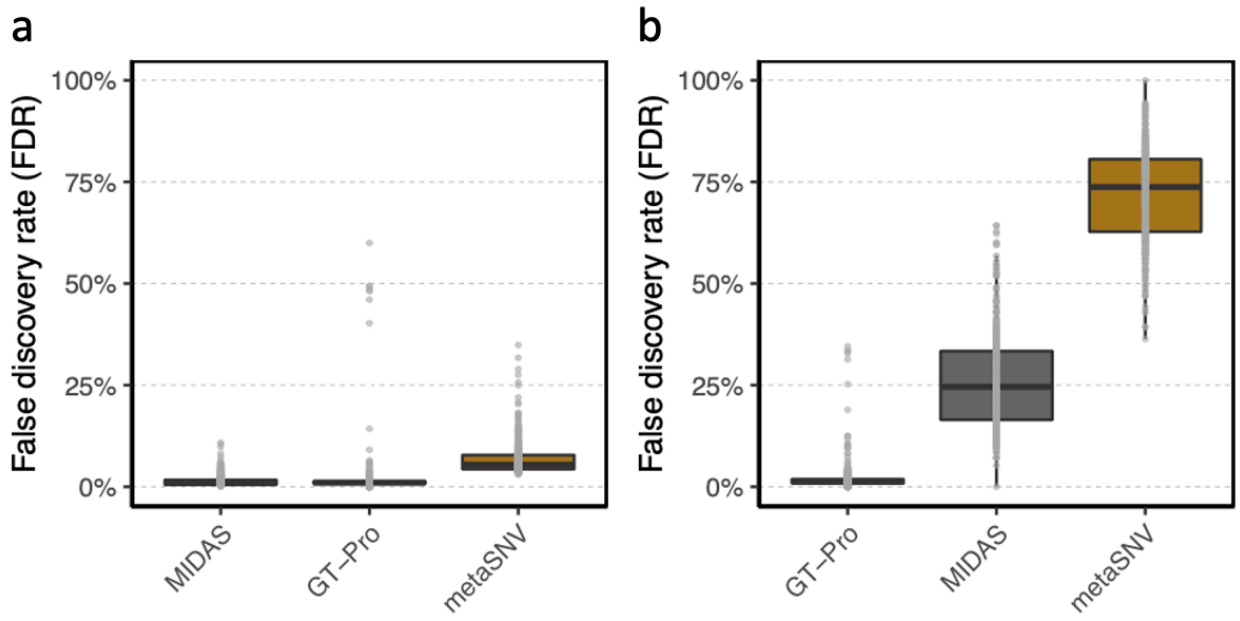


Figure S28
Sequencing coverage estimation comparison of GT-Pro and alignment methods. Sequencing coverage estimated using read counts at genotyped reference and non-reference sites when the simulated coverage is <= 1x (a) and >= 1x (b). Each observation is the estimate from metagenomic reads simulated with sequencing error from a single isolate genome.

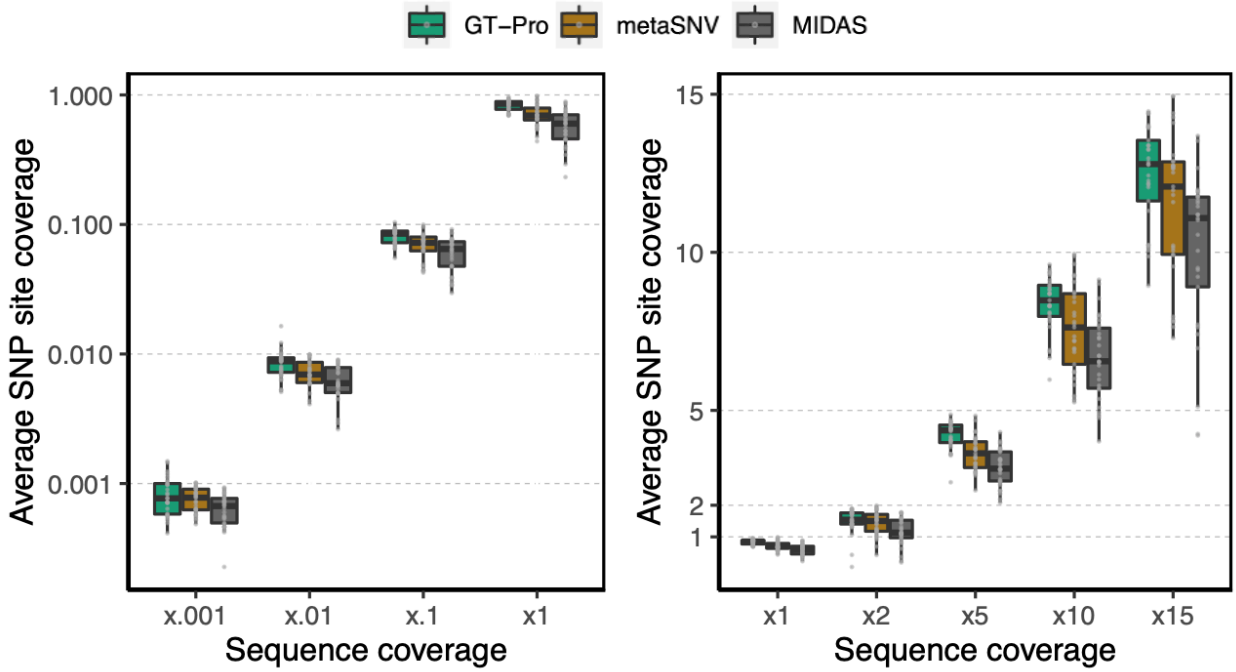


Figure S29
Default horizontal coverage cutoff selection using simulated reads. Metagenomic reads were simulated at a coverage of 15x from the same genomes and species as Fig 4a and b. The red dashed line is a horizontal coverage cutoff at 0.05, which we used to determine the presence of a species while limiting the number of false positive (FP) detections (see Methods).

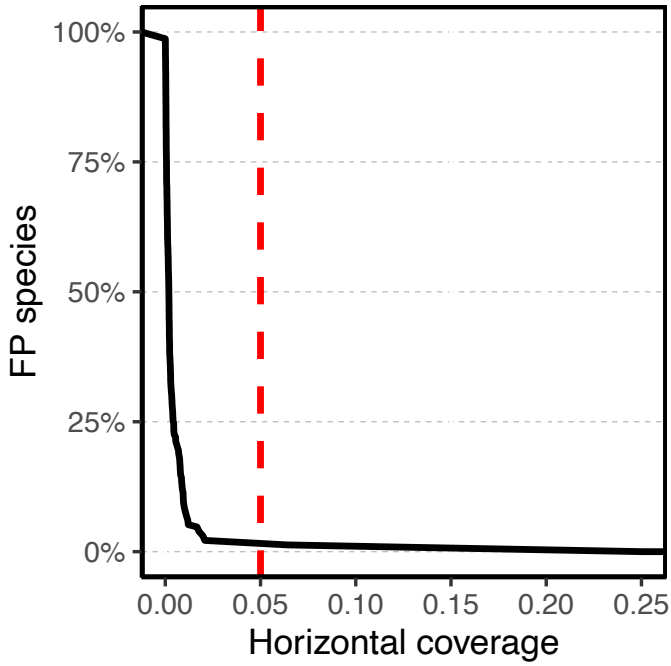


Figure S30
Amount of diversity captured in metagenotyping results on a Tanzanian cohort (n=40; Table S8) for GT-Pro, MIDAS, metaSNV and StrainPhlAn. Methods were run with their native databases, which contain different species, so results reflect the combination of what theoretically can be genotyped (species for each, SNPs per species for GT-Pro) and the sensitivity of each method to call genotypes across species with different coverages in the data. (a) number of detected species per sample. A horizontal coverage cutoff of 0.05 was applied to all methods except for StrainPhlAn. All species with non-zero relative abundance were counted for StrainPhlAn. (b) Median number of genotyped sites across detected species in each sample. A 1x coverage cutoff was applied to GT-Pro. 5x and 10x coverage cutoffs were applied to both MIDAS and metaSNV. StrainPhlAn was not included because it does not report individual SNPs.

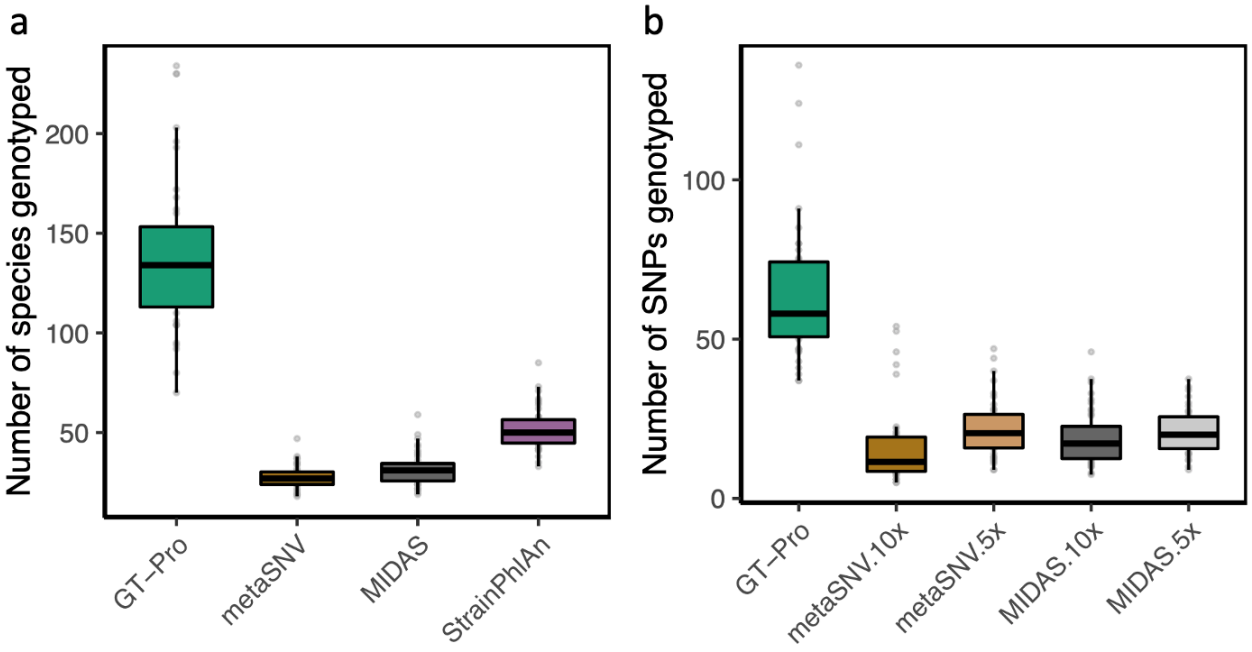


Figure S31

Agreement of metagenotyping output between GT-Pro and alignment methods (metaSNV, MIDAS) in samples from the Human Microbiome Project (HMP, above; Table S10) and Madagascar cohort (MDG, below; Table S11). Agreement was measured with both Jaccard distances (left) and correlation of allele frequencies (right) at all genotyped sites per sample. The color of stack bars indicates metaSNV (brown) and MIDAS (grey).

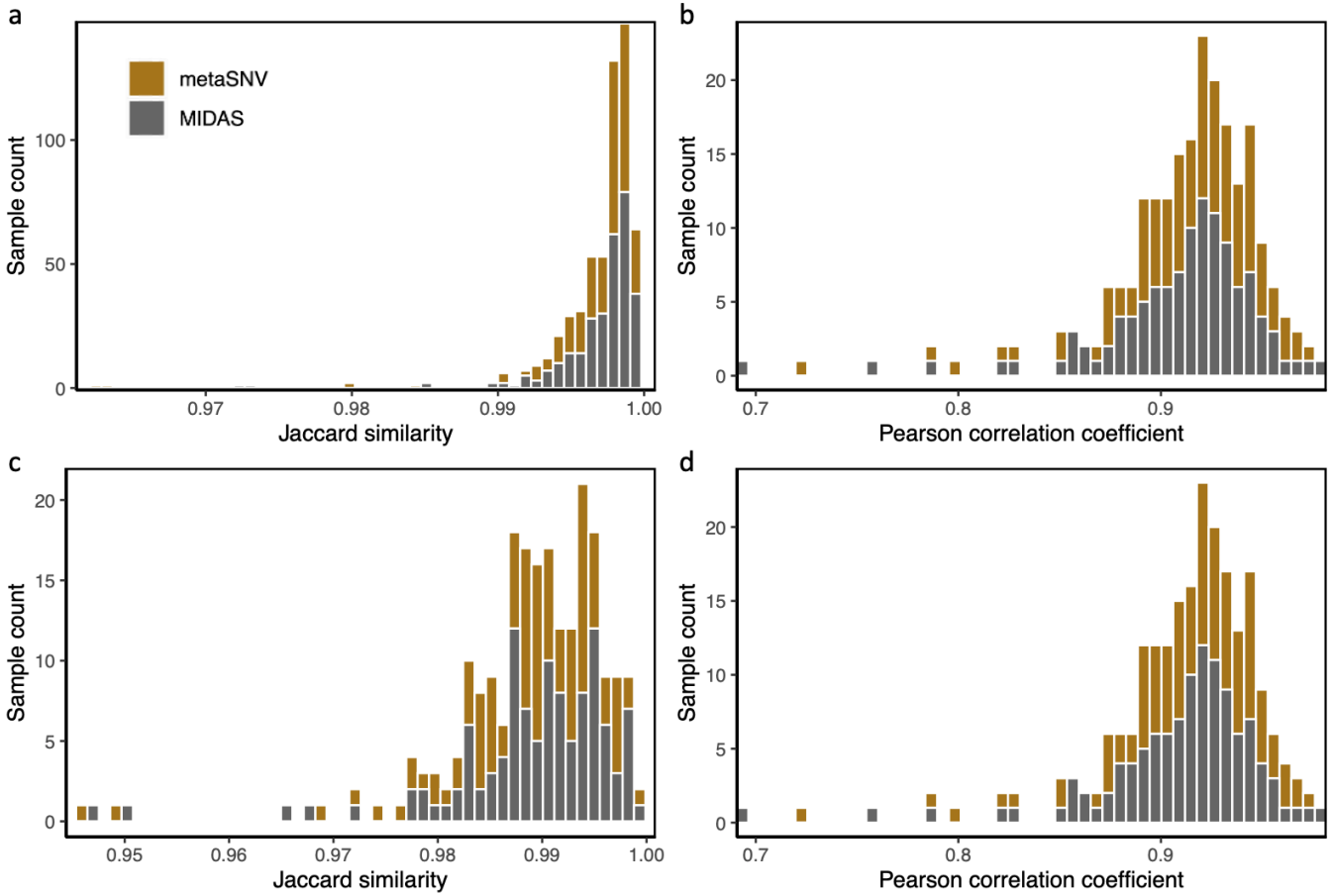


Figure S32
Paired genetic distance between samples from a from a North American IBD cohort (n=220; Table S12) based on different subsets of SNPs (left to right) for a given species (Intestinimonas butyriciproducens; species id: 102340). Left: all SNPs genotyped with default parameters (with metaSNV and MIDAS, includes rare SNPs), Middle: all common SNPs genotyped by GT-Pro, Right: all SNPs filtered and concatenated by StrainPhlAn.

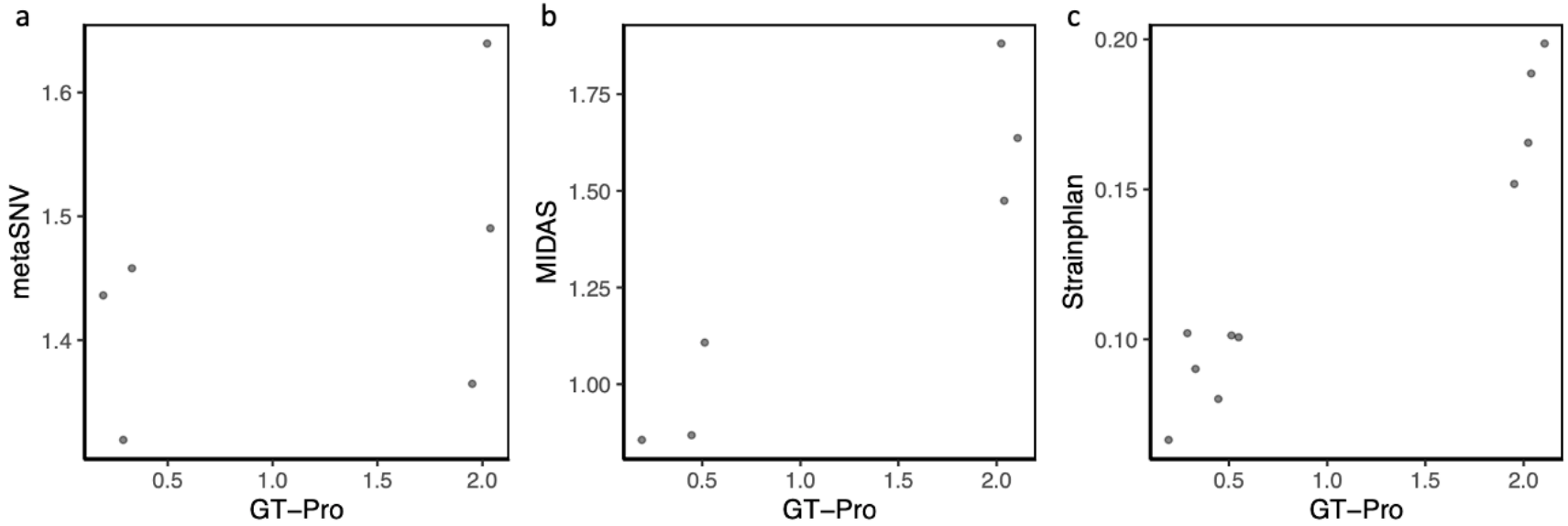


Figure S33

Pathogenic genes and GT-Pro SNPs in their flanking regions on contigs of the representative genome of *C. difficile* from GT-Pro database. SNPs were used to predict presence/absence of the pathogenic gene sets. Top: gene locations (colors: genic regions), bottom: flanking SNP locations (colors: genic or intergenic labels). (a) CdtLoc. (b) PaLoc, which spans two contigs.

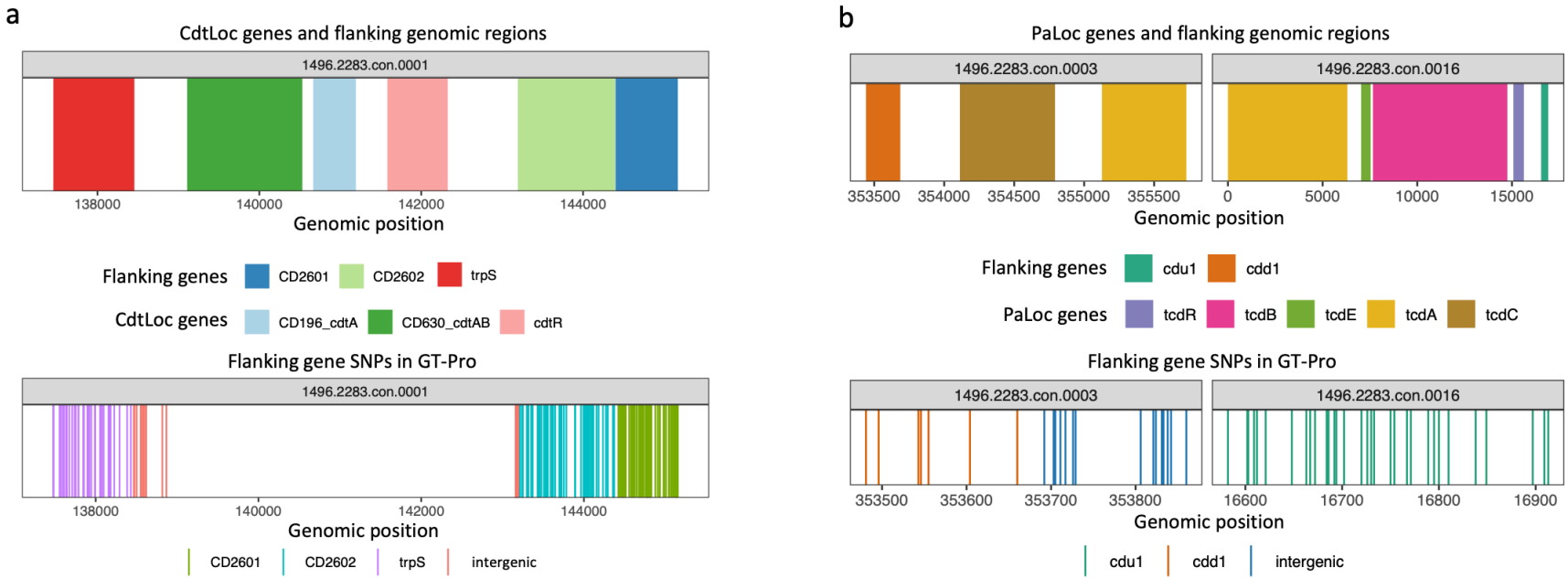


Figure S34

Prediction of presence/absence of pathogenic gene sets in the genomes 110 *C. difficile* isolates (Table S15) with a random forest classifier built using GT-Pro SNPs from flanking regions and 10-fold cross validation. Rows: SNPs (sidebar colors indicate genic regions), columns: genomes (sidebar colors indicate pathogenic gene set presence/absence in the genomes (observed) and according to the random forest (predicted). Heatmap cells show SNP allele: reference (green), alternative (purple), not detected (light grey). The random forest yielded a perfect prediction from the cross validation. SNPs are ordered by genomic coordinate. (a) CdtLoc and (b) PaLoc.

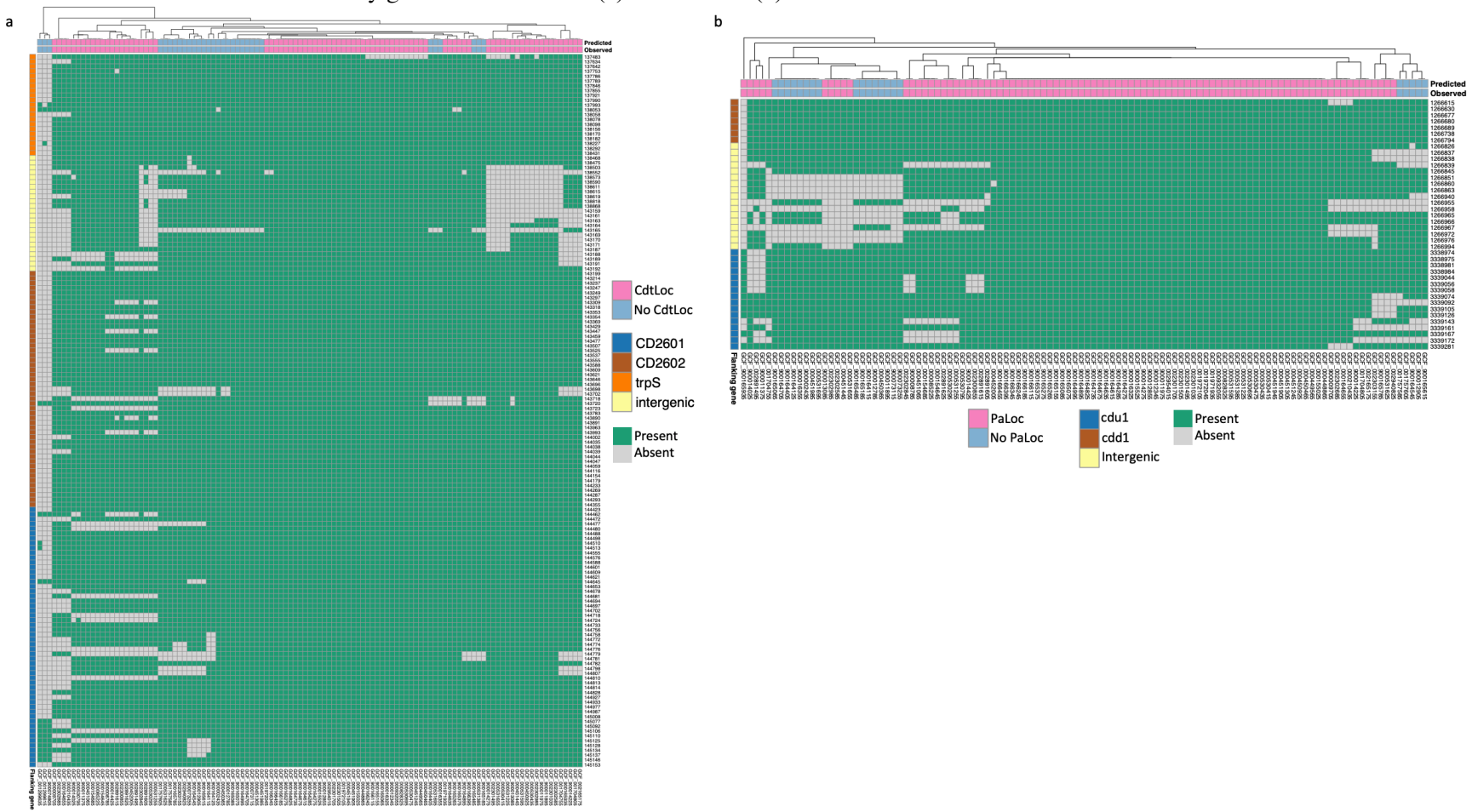


Figure S35

Random forest model is more confident of the presence of the entire pathogenic gene set when more of the genes are detectable in the metagenome. Plotted values for each metagenome are the random forest score (predicted probability of the entire pathogenicity gene set being present based on flanking SNPs) and the number of genes with >50% of length covered at least 1 read in the metagenome. Dotted line based on linear regression model. The level of correlation is indicated by Pearson's *r* and p-value. (a) CdtLoc and (b) PaLoc.

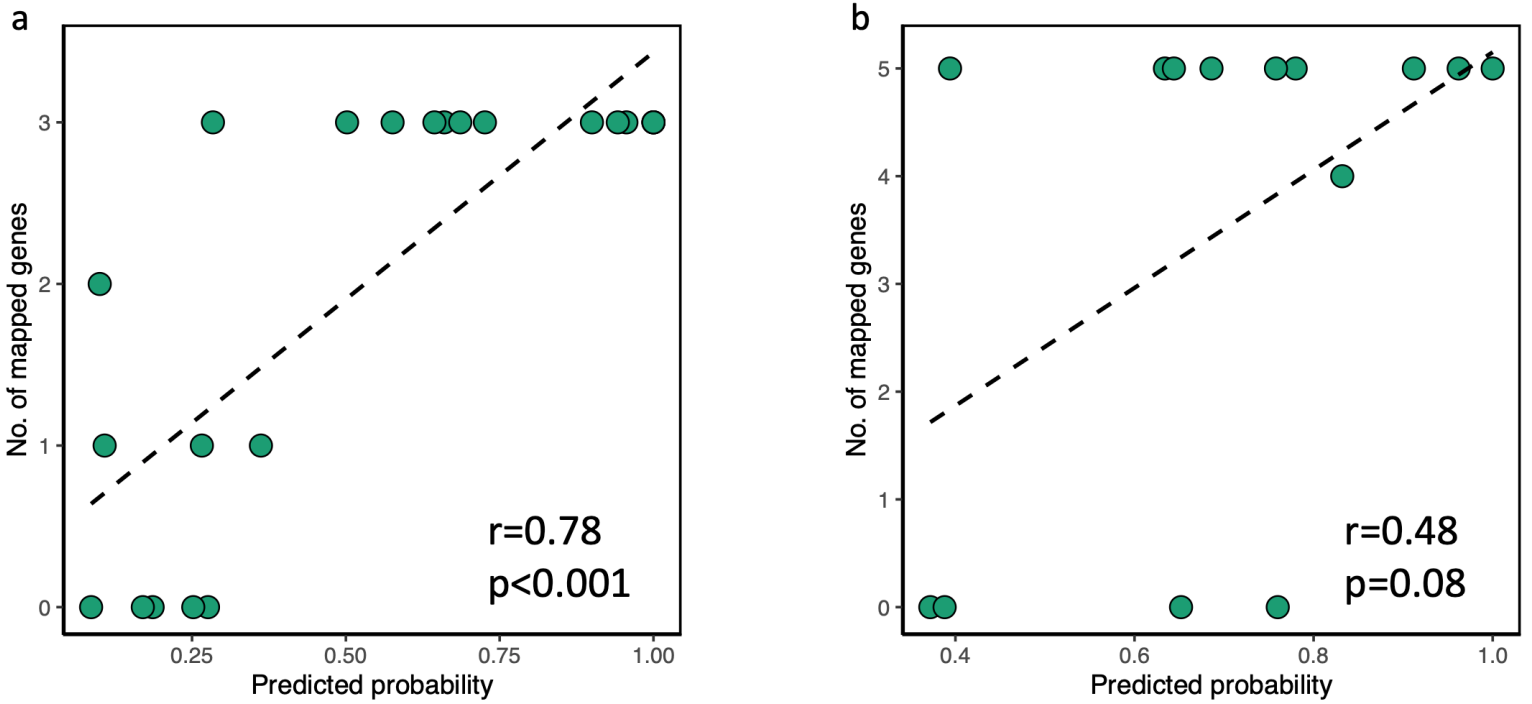


Figure S36
Visualization of metagenomic samples with novel subspecies from a North American IBD cohort (n=220; Table S12). Both diseased (purple) and control (blue) metagenomes are visualized in two dimensions alongside the UHGG genomes (green). Each plot is the result of applying UMAP to a matrix of 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. Presumptive novel subspecies clusters in (a) *Dialister invisus* (species id: 104158) and (b) *Dorea scindens* (species id: 101303) are boxed with grey lines.

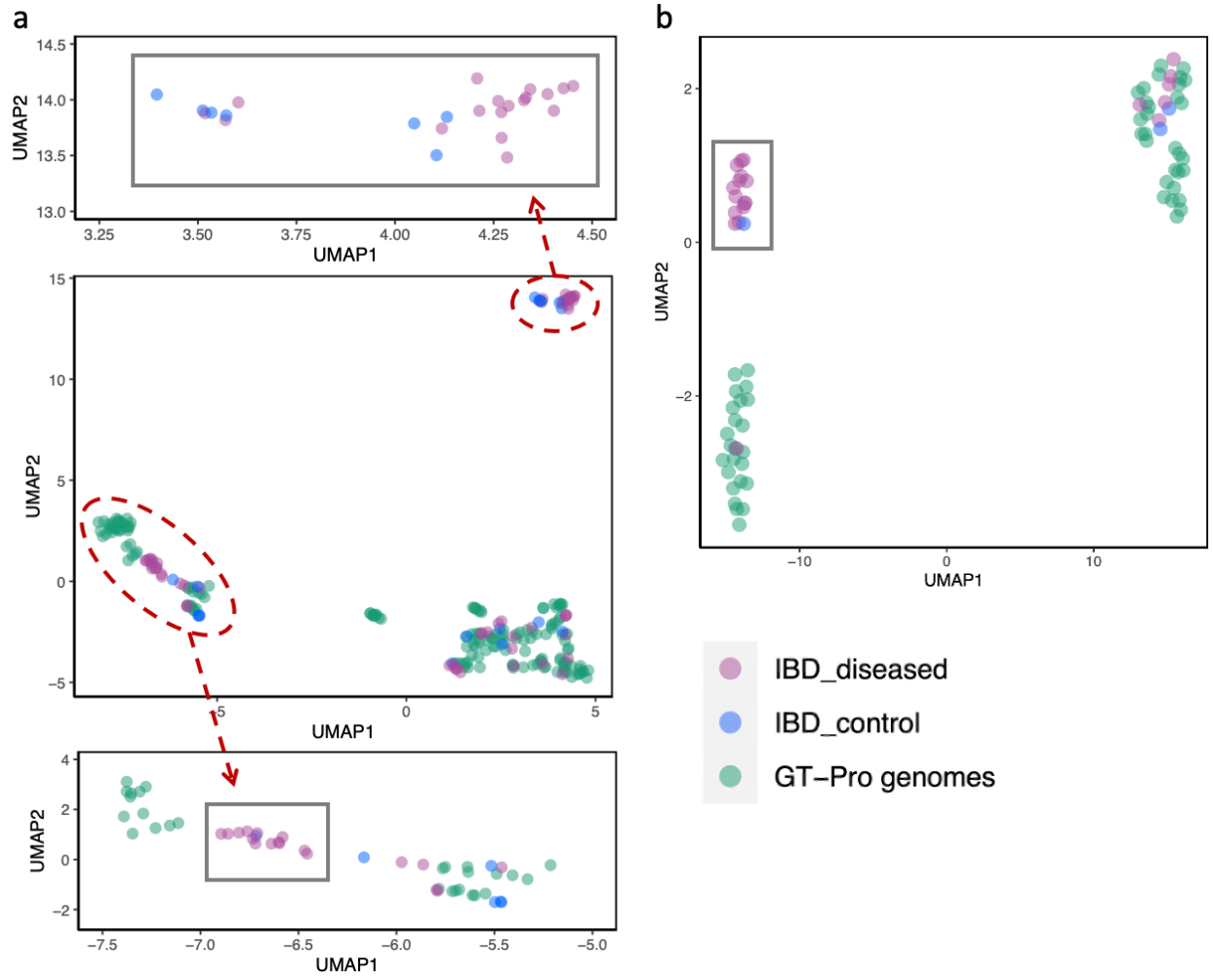


Figure S37
Genetic distance estimate comparison of GT-Pro and other methods (metaSNV, MIDAS and StrainPhlAn). Sequencing reads were simulated at 15x coverage from a total of 247 genomes from an arbitrary species (ID: 100113) including four CGR genomes. The reads were then spiked into one arbitrary metagenome (sample accession #: SRR6468562) from Franzosa et al. cohort per genome. A 5x coverage cutoff was applied to both MIDAS and metaSNV as suggested in the metaSNV paper. No coverage cutoff weas applied to GT-Pro and StrainPhlAn. Each dot represents a sample pair with genetic distance estimated from consensus allele sequences using RAxML. A dot is colored in red if the corresponding sample pair contains reads simulated from a CGR genome.

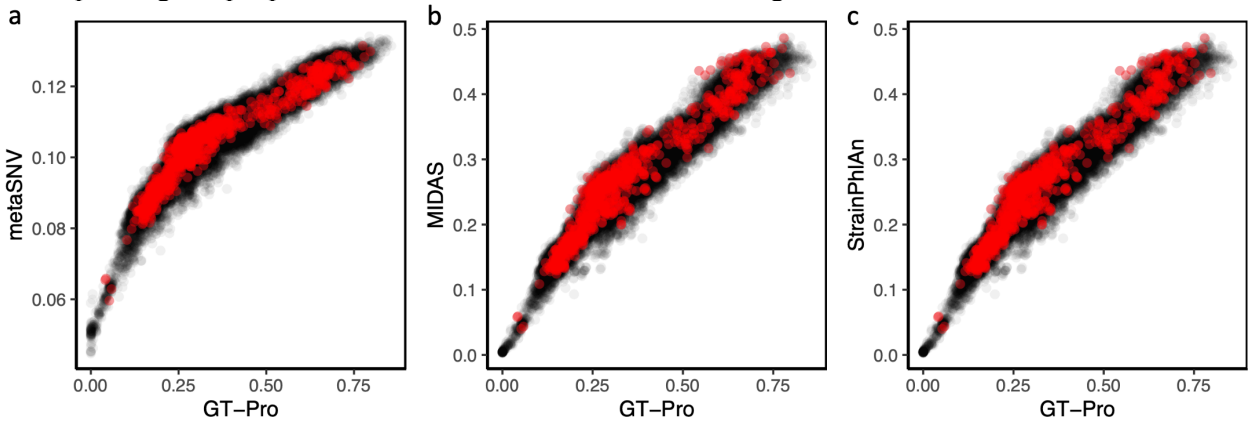


Figure S38
SNP recovery decay of GT-Pro, metaSNV and MIDAS. Sequencing reads were simulated at 15x coverage from a total of 24 genomes from an arbitrary species (ID: 102295) and spiked into one arbitrary metagenome (sample accession #: SRR6468551) separately. Ground truth SNPs were identified by aligning individual genomes to a reference genome using MUMmer. Each dot represents a sample comparing ground truth with correctly genotyped SNPs for a method.

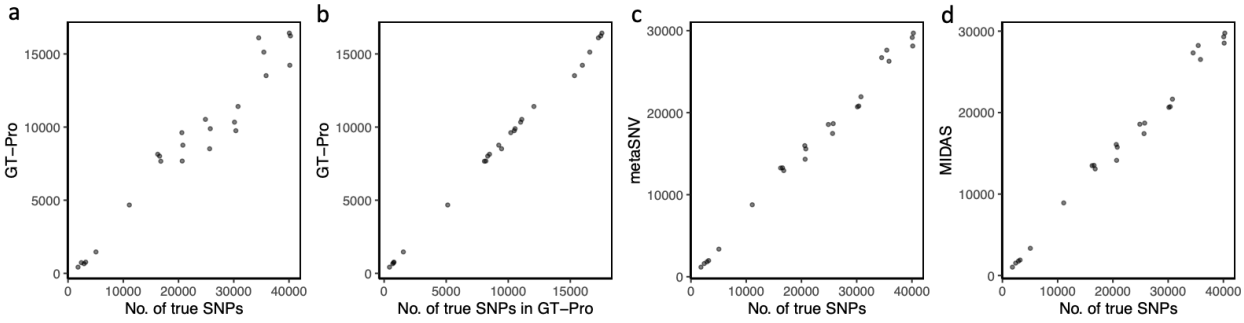


Figure S39
Allele sharing increases with relatedness of metagenomes. Scores calculated between pairs from a set of randomly selected samples (Table S14) from different hosts (inter-individual) and different samples from the same host (intra-individual), and technical replicates from the same sample.

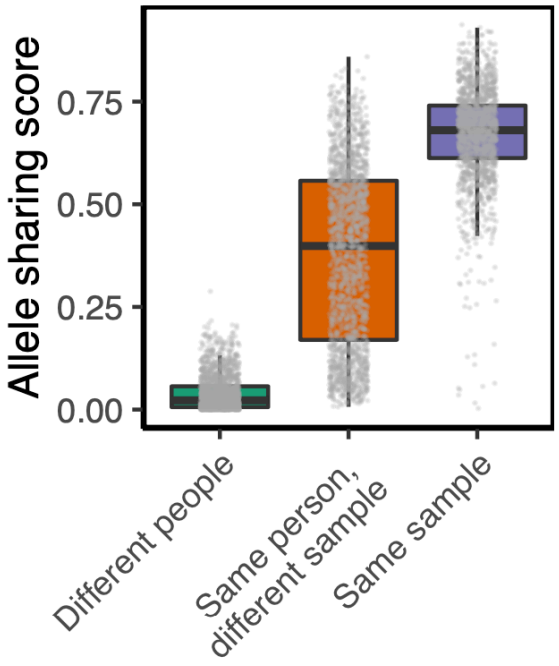


Figure S40

(a and b) Average allele sharing increases with geographic proximity. Scores calculated between pairs of samples across cohorts from different countries within the same continent (a) and from different continents (b). Continents: Africa (AF), Asia (AS), Europe (EU), North America (NA), Oceania (OC), South America (SA).

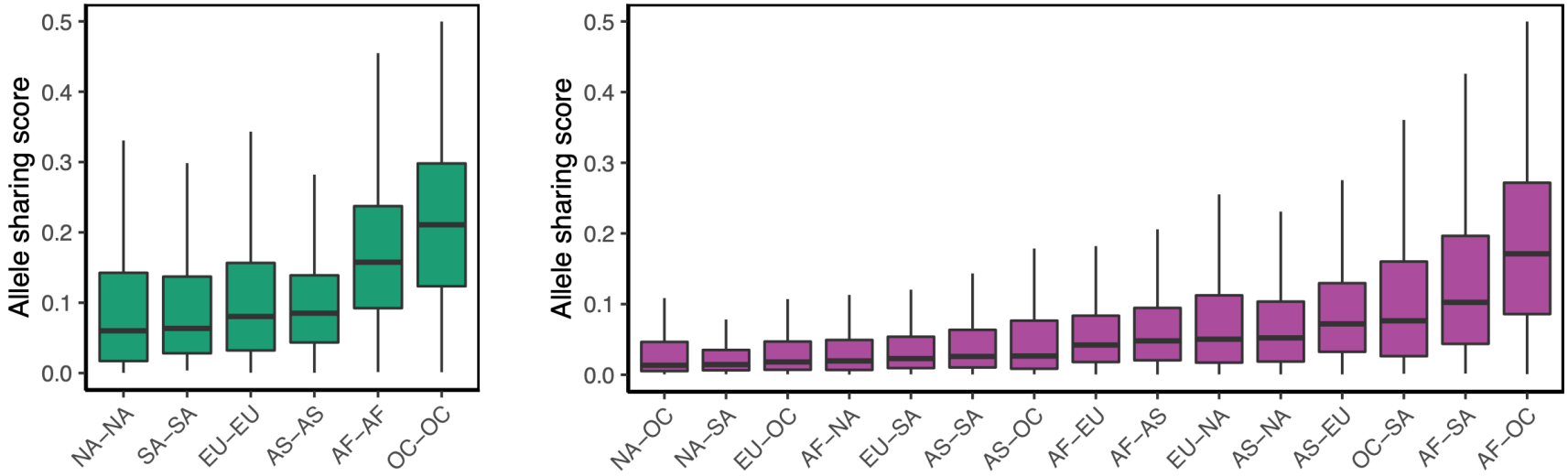


Figure S41
Comparison of allele sharing scores between pairs of samples from (a) same versus different continent and (b) same versus different country. *** p-value < 0.001 by two-sided Wilcoxon rank sum test with continuity correction.

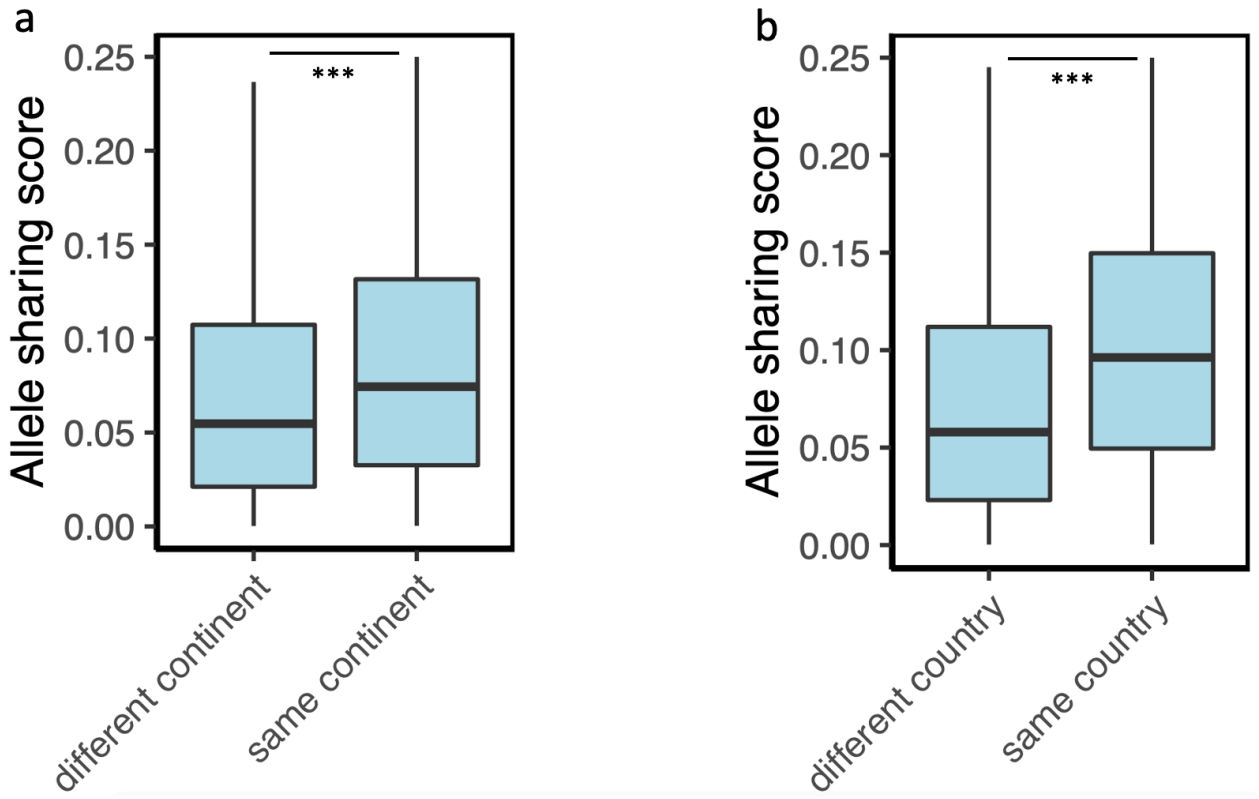


Figure S42

Pairwise allele sharing scores by individual species between samples from (a) the same continent and (b) different continents. Each point represents a mean allele sharing score for a species. Black horizontal line represents pairwise allele sharing scores using all species. The species are colored by their phyla.