# Enhanced metagenomics-enabled transmission inference with TRACS

**Gerry Tonkin-Hill[1,2,3,4], Yan Shao[5], Alexander E. Zarebski[6], Sudaraka Mallawaarachchi[1,2,4], Ouli Xie[7,8], Tommi Mäklin[9], Harry A. Thorpe[4], Mark R. Davies[3], Stephen D. Bentley[5], Trevor D. Lawley[5], and Jukka Corander[4,5,9]**

[1]Peter MacCallum Cancer Centre, Australia
[2]Sir Peter MacCallum Department of Oncology, The University of Melbourne, Australia
[3]Department of Microbiology and Immunology, The University of Melbourne, at the Peter Doherty Institute for Infection and Immunity, Australia
[4]Department of Biostatistics, University of Oslo, Norway
[5]Parasites and Microbes, Wellcome Sanger Institute, United Kingdom
[6]School of Mathematics and Statistics, The University of Melbourne, Australia
[7]Department of Infectious Diseases, The University of Melbourne at the Peter Doherty Institute for Infection and Immunity, Melbourne, Australia
[8]Monash Infectious Diseases, Monash Health, Melbourne, Australia
[9]Helsinki Institute for Information Technology HIIT, Department of Mathematics and Statistics, University of Helsinki, Finland

**Coexisting strains of the same species within the human microbiota pose a substantial challenge to inferring the host-to-host transmission of both pathogenic and commensal microbes. Here, we present TRACS, a highly accurate algorithm for estimating genetic distances between strains at the level of individual SNPs, which is robust to intra-species diversity within the host. Analysis of well-characterised Faecal Microbiota Transplantation datasets, along with extensive simulations, demonstrates that TRACS substantially outperforms existing strain aware transmission inference methods. We use TRACS to infer transmission networks in patients colonised with multiple strains, including SARS-CoV-2 amplicon sequencing data from UK hospitals, deep population sequencing data of *Streptococcus pneumoniae* and single-cell genome sequencing data from malaria patients infected with *Plasmodium falciparum*. Applying TRACS to gut metagenomic samples from a large cohort of 176 mothers and 1,288 infants born in UK hospitals revealed species-specific transmission rates between mothers and their infants. Notably, TRACS identified increased persistence of *Bifidobacterium breve* in infants, a finding missed by previous analyses due to the presence of multiple strains.**

**pangenome | microbial genomics | horizontal gene transfer**
**Correspondence:** ***gtonkin@unimelb.edu.au***

## Introduction

Host-to-host transmission is a fundamental process that shapes the interactions between humans and microbes. Tracking the spread of pathogens using genomics has become a major tool in public health, helping to prevent the spread of disease at both local and global scales. Beyond pathogens, understanding the transmission and colonisation dynamics of commensal microbes would greatly improve our understanding of microbiome assembly and maintenance, and how this is influenced by diet, lifestyle, culture, clinical interventions and social interactions. In addition, the human microbiome contains microbes with therapeutic potential to treat a variety of humans' diseases. The ability to identify candidate strains during clinical interventions such as Faecal Microbiota Transplantation (FMTs) and Live Biotherapeutic products would greatly de-risk and accelerate the development of microbiome-based therapeutics.

Whole genome sequencing (WGS) has reshaped our ability to infer transmission chains by enabling the detection of single-nucleotide variants, facilitating precise tracking of slowly-evolving pathogens such as Methicillin-resistance *Staphylococcus aureus* (MRSA)(1). However, current WGS approaches often focus on a representative strain genome of a single species, overlooking the microbial strain diversity within individuals. Recent efforts have begun to address this limitation using metagenomics, enabling the simultaneous consideration of multiple species and their strains in a single analysis (2, 3). Deep population sequencing of specific species can be achieved through targeted enrichment techniques like culture or PCR amplicon sequencing, allowing for fine-scale analyses of multiple co-colonising strains (4–6). Existing gold-standard methods for tracking transmission using metagenomic or deep population sequencing data have typically been developed for curated academic studies (3, 7–9). While these methods offer valuable insights, such as within-host variant calling and identification of selection pressures via dN/dS, they lack the speed and flexibility required for routine public health monitoring. Notably, they often lack the temporal resolution needed to accurately distinguish between strains transmitted recently (weeks to months) and distantly related genomes (years).

Tools that rely on reference marker gene databases including MIDAS (9, 10) and StrainPhlAn (7) consider only a small portion of a species genome (10-200 genes(2)) and do not attempt to separate within species diversity. This substantially limits the temporal resolution at which transmission can be inferred. Another common approach relies on de novo metagenomic assembly, including the inStrain (assembly mode)(3) and STRONG(11) pipelines. Assembly requires high sequencing coverage of the transmitted genome(8) and the large computational complexity prohibits its use in routine public health genomic surveillance. An alternative approach, used by StrainGE (8) and inStrain (database mode)(3) involves identifying the species found across the entire dataset and building a dataset-specific reference genome database for read alignment. This approach relies heavily on the similarity between the transmitted and reference genome and does not allow for the continuous integration of new samples into an analysis, making it unsuitable

for routine genomic surveillance. Moreover, recombination and shared homology between strains of the same species within a sample is generally not accounted for in existing algorithms, which can have a major impact on the accuracy of transmission inference when considering metagenomic and population sequencing data. To address these issues, we developed TRAnsmision Clustering of Strains (TRACS), a highly accurate and easy-to-use algorithm for establishing whether two samples are plausibly related by a recent transmission event.

The TRACS algorithm distinguishes the transmission of closely related strains by identifying genetic differences as small as a few Single Nucleotide Polymorphisms (SNP)s, which is crucial when considering slow-evolving pathogens. Critically, unlike existing tools, TRACS was designed to estimate a *lower bound* of the SNP distance and can incorporate sampling date information. This makes TRACS ideal for definitively ruling out putative transmission events, an essential requirement for public health applications. Using novel statistical filtering techniques, TRACS controls for major sources of error including variable sequencing coverage, within-species recombination and sequencing errors.

We demonstrate using comprehensive simulations, and by considering well characterised Faecal Microbiota Transplant (FMT) metagenomic datasets that TRACS has superior performance to existing metagenomic transmission inference methods and can reliably identify plausible transmission events. We apply TRACS to a diverse set of pathogens including viruses, bacteria and parasites, demonstrating the scalability and versatility of the algorithm. Finally we consider the development of the infant gut microbiome and species specific transmission rates between mothers and infants using a large cohort of 1,464 mothers and infants.

## Results

### TRACS addresses key challenges in metagenomic transmission inference.

The TRACS algorithm consists of a modular pipeline for tracking multiple species and their strains using either metagenomic, multi-strain deep population sequencing data of individual species within a sample or individual colony sequencing. The algorithm allows for the efficient addition of new samples to an analysis, a key requirement for near real-time use in public health settings, such as emerging outbreaks.

The initial 'alignment' stage of TRACS uses the hash-based search algorithm 'Sourmash' to identify a set of reference genomes that best represent the species and their strains in a sample(12). Unlike alternative approaches, TRACS avoids the error prone deconvolution of reads into different genome bins and instead aligns the entire read set to each reference genome individually, generating a count of the alleles observed at each site(13).

A novel set of statistical filtering algorithms are then applied to exclude regions affected by recent horizontal gene transfer, poor alignment quality, and low sequencing coverage (Figure 1). This includes a scan statistic, similar to those used to identify recombination in phylogenetics(14), and an empirical Bayes approach to account for regions of a reference genome with insufficient coverage to accurately represent multiple strains present within a sample (described in Methods). The resulting filtered alignments are converted into a Multiple Sequence Alignment (MSA) for each reference genome, allowing for polymorphic sites using IUPAC ambiguity codes. TRACS incorporates a fast, IUPAC-aware pairwise SNP distance estimation algorithm, enabling rapid inference of genetic distances.

To account for the sampling dates, an updated version of the TransCluster algorithm(15) is included, which estimates the expected number of intermediate hosts separating two samples. The distance estimated by this method is straightforward to interpret, facilitating the incorporation of both genetic and temporal distances between samples. Using this approach, transmission clusters are inferred using the single linkage hierarchical clustering algorithm. While the alignment component of TRACS is designed for metagenomic and population sequencing data, MSAs produced by alternative tools designed for isolate-data, such as Snippy(16), can also be directly input into the transmission distance algorithm.

A key challenge in using pairwise distance methods to infer transmission chains is determining an appropriate threshold for when two strains are likely connected by recent transmission. TRACS introduces a novel method that uses a mixture distribution to distinguish between recent transmission events and strains that are distantly related by years of evolution. This results in thresholds that are more specific than those provided by previous methods such as those that rely on Youden's index (17). Similar to previous efforts (2), this approach leverages data from known closely related samples, such as strains from the same individual, and distinguishes them from distantly related samples.

The TRACS algorithm is implemented in python and C++, and is available under the MIT open source licence at https://github.com/gtonkinhill/tracs/.

### TRACS achieves superior performance on simulated transmission pairs.

To assess the ability of TRACS to accurately identify transmitted strains in pairs of samples containing multiple strains of the same species, we simulated WGS reads from mixtures of common *Streptococcus pneumoniae* strains. In each pair we selected one species to be transmitted at a specified SNP distance (5, 50 and 500 SNPs). The TRACS algorithm was compared to InStrain, StrainGE, and StrainPhlAn.

All algorithms except TRACS substantially overestimated the genetic distances (Figure 2A). The bias in these estimates prevents those algorithms from reliably excluding the possibility of recent transmission. StrainGE was the best-performing algorithm after TRACS, likely due to its use of competitive mapping between genomes within a species, which helps control for the effects of horizontal gene transfer (HGT) and shared homology. However, StrainGE only reported relatively accurate SNP distances at higher values (500 SNPs), which substantially exceed the typical thresholds used to distinguish recent transmissions in most bacterial

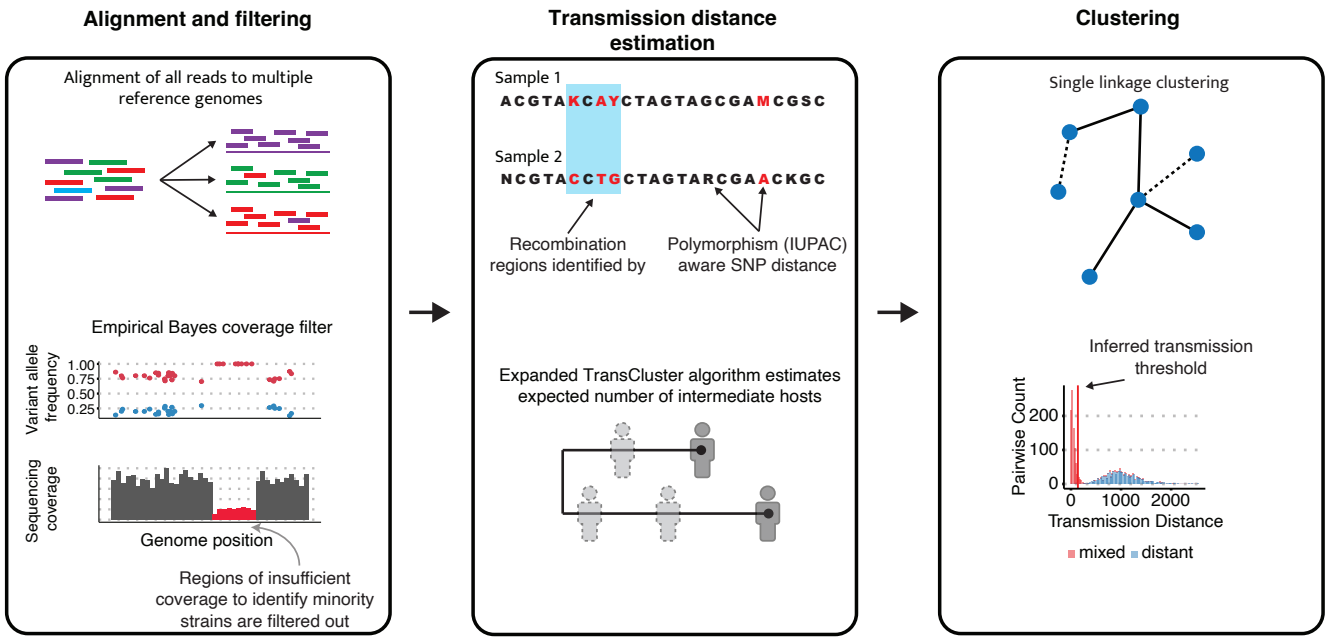



**Fig. 1.** A schematic illustrating the key components of the TRACS algorithm. Left: reads are aligned to each reference genome identified by Sourmash separately. An empirical Bayes method is used to pinpoint genome regions with insufficient coverage for minority strain identification. Centre: alignments from each sample are transformed into a multiple sequence alignment using IUPAC ambiguity codes to represent multiple alleles at a single site. Rapid pairwise SNP distances are then calculated, excluding potential recombination regions by identifying areas with high SNP density. The TransCluster algorithm can optionally be applied to estimate the expected number of intermediate hosts between two samples. Right: the resulting pairwise transmission distance estimates are clustered using single linkage hierarchical clustering to infer putative transmission clusters. Transmission distance thresholds are inferred using a mixture distribution to separate sample pairs that are known to be distantly related from those that include recent transmissions.

pathogens(18). StrainPhlAn considers the dominant lineage in each sample, limiting its sensitivity in cases with multiple strains of the same species. In addition, we found that TRACS maintained a similar level of accuracy when applied to simulated Nanopore R10.4.1 reads, demonstrating its applicability to various high-accuracy sequencing technologies (Supplementary Figure 2).

To confirm these results with real sequencing data, we considered 13 laboratory mixtures of different *S. pneumoniae* strains (19). In each pair of samples, there was at least one identical pneumococcal genome. Therefore, each algorithm should infer zero SNPs. As shown in Figure 2B, TRACS was the only algorithm that reliably inferred zero SNPs in all cases. StrainPhlAn correctly identified zero SNPs in some cases; however, its reliance on marker genes can lead it to systematically underestimate SNP distances, as shown in the subsequent analysis of FMT data.

To further explore the reconstruction of transmission chains from metagenomic data involving multiple distinct species, we simulated mixtures of reference genomes representing common gut bacteria (as described in Methods). In these simulations, a single species identified in a previous study of transmission using metagenomics was simulated to transmit with a specified SNP distance (20). TRACS remained the best performing algorithm under these settings (Figure 2B). InStrain overestimated the SNP distance for most simulated pairs due to genetic differences between the transmitted and reference genomes in its database. When the genome of the transmitted strain closely matched the reference, InStrain provided accurate results, as shown by the outlying points in Figure 2C. In contrast, the performance of StrainGE and StrainPhlAn improved consistently when only a single strain per species was present. This improvement is driven by the reduced rates of recent horizontal gene transfer and shared homology among the simulated genome mixtures. Despite the improvement, StrainPhlAn consistently underestimated genetic distances between very closely related genomes (5 SNPs), indicating it would not be suitable for distinguishing recent transmission in slow-evolving species such as *M. tuberculosis* and *S. pneumoniae*.

### Enhanced estimates of pathogen transmission across diverse taxonomic groups in high-burden settings.

A major advantage of the TRACS algorithm is its ability to identify putative transmission events in cases where the hosts are colonised with multiple strains of the same species. This is a major concern in many high disease burden settings. Unlike alternative algorithms such as StrainPhlAn, the TRACS algorithm is applicable across diverse taxonomic groups including parasitic, viral and bacterial species. To demonstrate the effectiveness of the TRACS algorithm across a wide range of pathogens we considered well characterised datasets across three different settings.

#### *SARS-CoV-2.*

Although infection with multiple SARS-CoV-2 strains is relatively rare (21), the rates of occurrence can be higher in locations with high disease burdens, such as within hospital wards, particularly when infection control procedures have broken down. Moreover, in the early stages of an outbreak of a novel pathogen, the impact of sequencing errors and hyper-

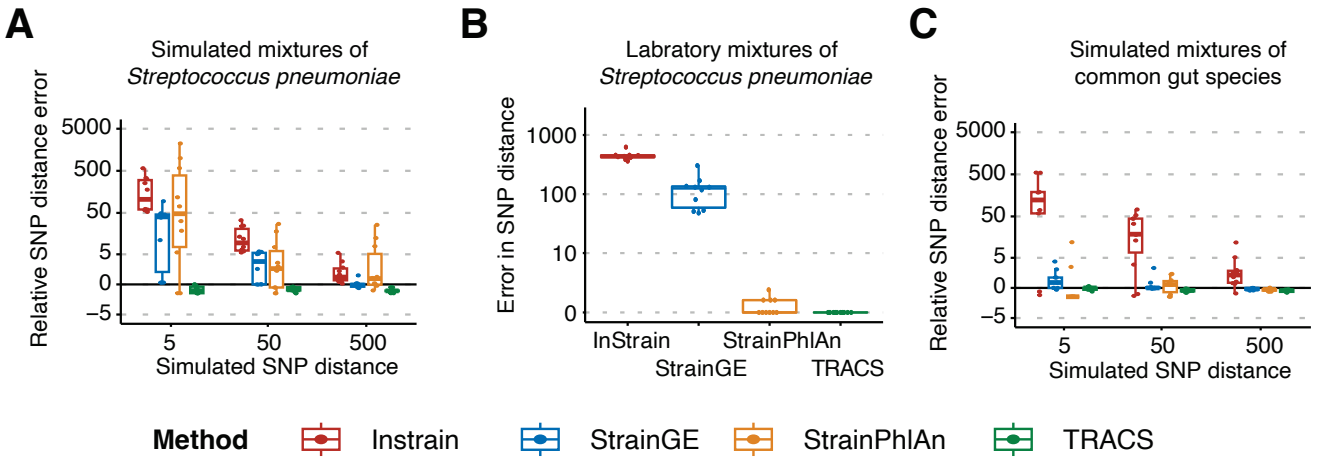



**Fig. 2. Accuracy of SNP distance estimation algorithms on simulated and laboratory mixtures.** A) The relative error in the estimated pairwise SNP distance across four algorithms, when applied to a simulated mixture of pneumococcal genomes. A single genome is simulated to have been transmitted between each pair of samples with a SNP distance given on the x-axis. Ten replicate simulations were performed for each parameter set with boxplots indicating the distribution in relative error rates. Points closest to zero are the best performing, with positive values indicating an overestimation of the SNP distance and negative values indicating an underestimation. B) The results of running each algorithm on artificial laboratory mixtures of pneumococcal strains from Knight et al.(19). Each pair of samples included at least one identical strain and thus the correct SNP distance should be zero. Except for TRACS, all algorithms erroneously estimate elevated SNP distances. C) The same simulation study as shown in panel A, but with simulations that include mixtures of genomes from common gut bacterial species. Only one genome per species was included in each sample.

mutating sites is typically unknown and can present a challenge to early efforts to track the spread of a pathogen (22).

To demonstrate the ability of the TRACS algorithm to account for these challenges, we considered 37 of 1181 SARS-CoV-2 samples, collected from the East of England in early 2020 which were processed using Illumina deep amplicon sequencing(23). These samples were found to contain multiple distinct strains after sequencing each sample in replicate to account for sequencing errors (24). These were compared to samples containing a single strain within the mixture. Thus, the smallest SNP distance between samples should be zero. To investigate the utility of applying TRACS to these samples, we compared the inferred SNP distance using TRACS to a consensus based approach, with and without filtering for problematic sites such as hypermutable loci and regions impacted by the ends of amplicon sequencing reads (Figure 3A) (22).

In cases where the shared strain was in the majority within the mixed sample, the filtered consensus approach correctly inferred zero SNPs. However, as expected, when the shared strain was in the minority, a consensus based approach consistently overestimated the SNP distance between samples despite filtering out problematic regions. In contrast, TRACS correctly inferred 0 SNPs between samples in all cases without requiring manual filtering of problematic sites. Figure 3B indicates the transmission network inferred assuming 0 SNPs between strains for those mixed samples with a geographic location. The dotted edges indicate transmission links that would have been missed using a standard consensus based framework. This example highlights the ability of TRACS to both account for the presence of multiple strains and to robustly control for the impacts of sequencing errors and hypermutable sites that frequently result in polymorphic variants within samples.

***Streptococcus pneumoniae.***

An added benefit of the TRACS algorithm is its ability to estimate the expected number of intermediate hosts between two samples. It achieves this by accounting for the sampling dates and SNP distances between samples, using an extended version of the TransCluster algorithm. Accounting for sampling dates is crucial for tracking slow-evolving bacterial pathogens such as *S. pneumoniae*.

To demonstrate this approach, we considered 3,761 nasopharyngeal swabs taken from 468 infants and 145 of their mothers living in a refugee camp in Thailand, where culture based enrichment and whole pneumococcal population Illumina deep sequencing had been performed as part of a previous study (5). TRACS was run on this dataset using a custom database of pneumococcal reference genomes from the Global Pneumococcal Sequencing project (25). The expected number of intermediate hosts between all pairs was compared to the geographic distance separating the households of participants (Figure 3C-D). Although the original TransCluster algorithm was previously applied to this dataset, it only estimates individual probabilities, such as the probability of direct transmission, without estimating the overall expected number of intermediate hosts.

There was a strong correlation between geographic distance and recent transmission, despite the small size of the refugee camp (2.4 $km^2$). Interestingly, this correlation was not found for all the major lineages. In particular, common multi-drug resistant lineages such as Global Pneumococcal Sequencing Cluster (GPSC) 1, had no clear association between transmission and geographic distance within the camp (Figure 3D). These lineages are known to have longer carriage durations, which can obscure geographic transmission signals (26, 27). Thus the observed signals are likely driven by lineages that transmit faster and exhibit shorter carriage durations, such as GPSC 20. Understanding the different dynamics governing

the transmission of these lineages is crucial for the design of interventions aimed at reducing pneumococcal disease.

***Plasmodium falciparum.***

The increased genome size of major parasitic pathogens has hampered the adoption of routine WGS in disease surveillance. However, the rapidly decreasing cost of sequencing is leading to the increasing use of WGS to track major parasite populations including the malaria parasite, *P. falciparum*, which causes over half a million deaths each year (28). Multiple strains of *P. falciparum* are frequently found within people living in endemic areas with high burdens of disease (29). To investigate the ability of the TRACS algorithm to accurately identify shared strains within mixed *P. falciparum* infections, we considered a dataset involving 49 samples from Chikh-wawa, Malawi that were positive for *P. falciparum* (30). Both Illumina bulk sequencing of mixed populations in addition to single cell and single clone enrichment had been performed as part of a previous study leading to 49 mixed WGS samples and 509 single genomes(30). Figures 3E and 3F indicate the distribution of the inferred SNP distances between single cell genomes and bulk samples from either the same sample or between samples. The clear separation of SNP distances between genomes sequenced from the same sample and between samples indicates that TRACS can accurately identify shared strains within this data. As WGS of malaria parasites becomes more routine in public health settings, the TRACS algorithm can be used to accurately determine if a pair of samples could be related by a recent transmission event. A fine scale understanding of transmission dynamics in high burden settings will aid in the design of strategies aimed at reducing and eventually eliminating malaria.

## Improved estimates of engraftment in faecal microbiota transplant triads.

Faecal microbiota transplantation (FMT) involves transferring a donor stool sample to a recipient, often via orally-delivered capsules, and has demonstrated patient benefit to treat infections, autoimmune diseases, graft versus host diseases and cancers. Development of microbiome drugs often requires the identification of specific bacterial strains with beneficial properties from complex microbiomes containing 100-1000s strains. To further evaluate the efficacy of the TRACS algorithm in metagenomics applications, we analysed an extensively characterised Faecal Microbiota Transplantation (FMT) dataset with well-defined transmission (or engraftment) relationships between samples(31). The capability of each algorithm to detect engraftment links between human gut microbiome samples from donors and recipients was examined by considering 23 previously published FMT triads(31). Metagenomic samples were originally collected from donors and recipients before and after transplantation across three patient cohorts, including individuals with *Clostridium difficile* infections, inflammatory bowel disease (IBD), and recurrent multi-drug resistant infections.

The TRACS algorithm was run on each of these samples and the results compared with StrainPhlAn, InStrain and StrainGE (Figure 4A-C). As StrainGE requires substantial computing resources to consider transmission of every observed genus we restricted this analysis to *Bifidobacteria*. The inferred number of engraftment events was then inferred, comparing samples from the same patient post-FMT and those from different cohorts (Figure 4A). Species specific SNP thresholds were chosen using both the novel mixture distribution method implemented in TRACS and the Youden index as used previously to consider the person-to-person transmission landscape of the gut and oral microbiomes (2) (Supplementary Figure 1). No strain sharing is expected between different cohorts, and thus inferred transmissions between cohorts are highly likely to be false positives. In contrast, we expect to see bacterial strains persist within a patient's gut microbiome over time. However, it is likely that these numbers are inflated by false positives at a rate similar to those observed in the between-cohort results.

StrainPhlAn had the highest false positive rate which increased substantially using the Youden method for selecting SNP thresholds (Figure 4A, Supplementary Figure 1). Consistent with our simulation results, InStrain was the most conservative method, performing well when there is high sequence similarity between the reference and transmitted genomes. However, InStrain frequently overestimated SNP distances in cases of greater sequence divergence from the reference or when multiple strains of the same species were involved, as indicated by our previous simulation based analysis. Using sample-specific de novo assembled reference genomes could improve the sensitivity of InStrain. However, this approach is impractical for many public health applications and large-scale transmission studies, so it is not considered here. In contrast to alternative algorithms, TRACS achieved a high sensitivity, with a low false positive rate.

To investigate this further, we considered the *Bifidobacteria* genus, and in particular the transmission of *Bifidobacterium longum*, a common pioneer coloniser of the infant gut microbiota(32, 33) which is often used as an infant probiotic. InStrain and StrainPhlAn were run using their default databases, while strainGE was supplied with a representative set of *Bifidobacteria* reference genomes following the user guide (Methods). TRACS was run using the GTDB reference database(34).

TRACS was the most sensitive algorithm at detecting the transmission of *B. longum*(Figure 4B). InStrain and StrainGE proved to be the most conservative algorithms, while StrainPhlAn identified the highest rate of sharing. This high rate of sharing likely includes many false positives due to StrainPhlAn's reliance on a small set of representative genes. Despite having *B. longum* genomes present in the reference database, StrainGE failed to identify this species amongst the engrafted strains. In contrast, InStrain identified *B. infantis* in its transplanted strains as *B. longum* and *B. infantis* are considered as one species in InStrain's default database (UHGG).

Unlike competing algorithms, TRACS consistently identified cases of engraftment involving multiple strains of *B. longum* present within a single sample. One example is given in Figure 4C, where the frequency of the engrafted strains was reversed in two patients who received the same donor

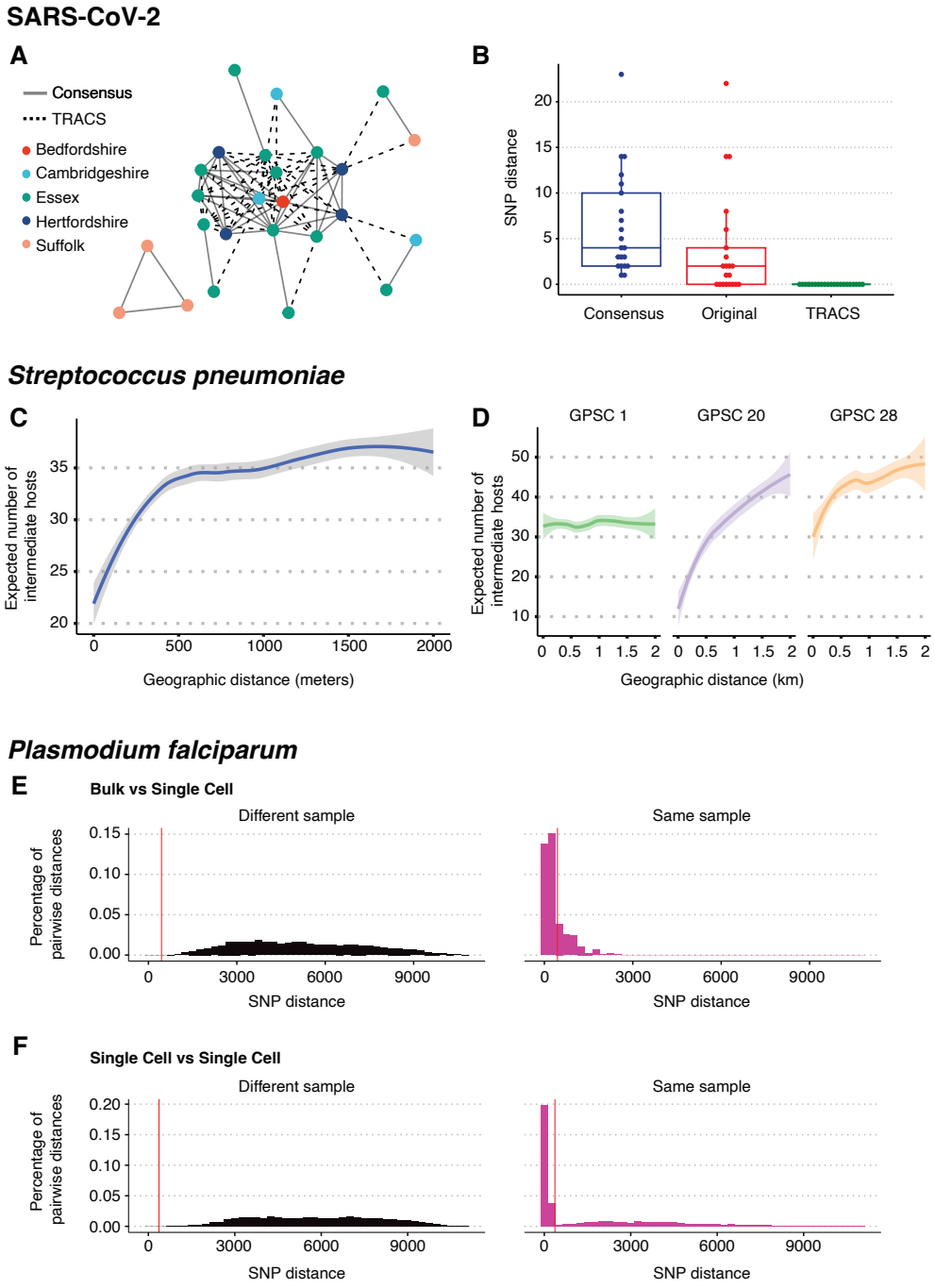



**Fig. 3. Transmission inference across diverse taxonomic groups using TRACS.** A) The inferred SNP distance between pairs of samples containing the same strain, where at least one sample includes multiple distinct strains. The consensus approach considers the dominant allele across the entire SARS-CoV-2 reference genome, whereas the 'original' approach excludes sites frequently filtered out to avoid hypervariable or error-prone regions. The TRACS algorithm correctly infers zero SNPs in all pairs. B) A SARS-CoV-2 transmission network comprising samples that contain multiple distinct strains. Solid lines indicate transmission events observable using TRACS in addition to typical consensus-based sequencing, while dotted lines represent additional links identified by TRACS alone. C) The expected number of intermediate hosts between each pair of pneumococcal samples taken from different individuals in the Maela refugee camp versus the geographic distance between their homes within the camp with a LOESS smoothing. Sample pairs with a divergence time outside the establishment of the camp in 1984 were excluded. D) Similar to C, but only points involved in the three most common global pneumococcal sequence clusters (GPSCs) are shown. A strong geographic signal is absent for the non multi-drug resistant lineage GPSC 1, which is known to have a longer carriage duration. E) and F.) The distribution of SNP distances inferred by TRACS for single cell vs. single cell, and bulk vs. single cell samples. The vertical red line indicates the SNP threshold inferred using the TRACS mixture distribution approach. The high number of short SNP distances between single cells from the same sample indicates that TRACS can accurately distinguish closely related genomes within mixed infections of *P. falciparum*.

stool. TRACS was the only algorithm to correctly identify that these patients shared similar strains. Identifying cryptic transmission in studies such as these is crucial for developing accurate models of microbiome colonisation dynamics, a key component of efforts to identify strains with health benefits to develop microbiome-based therapies.

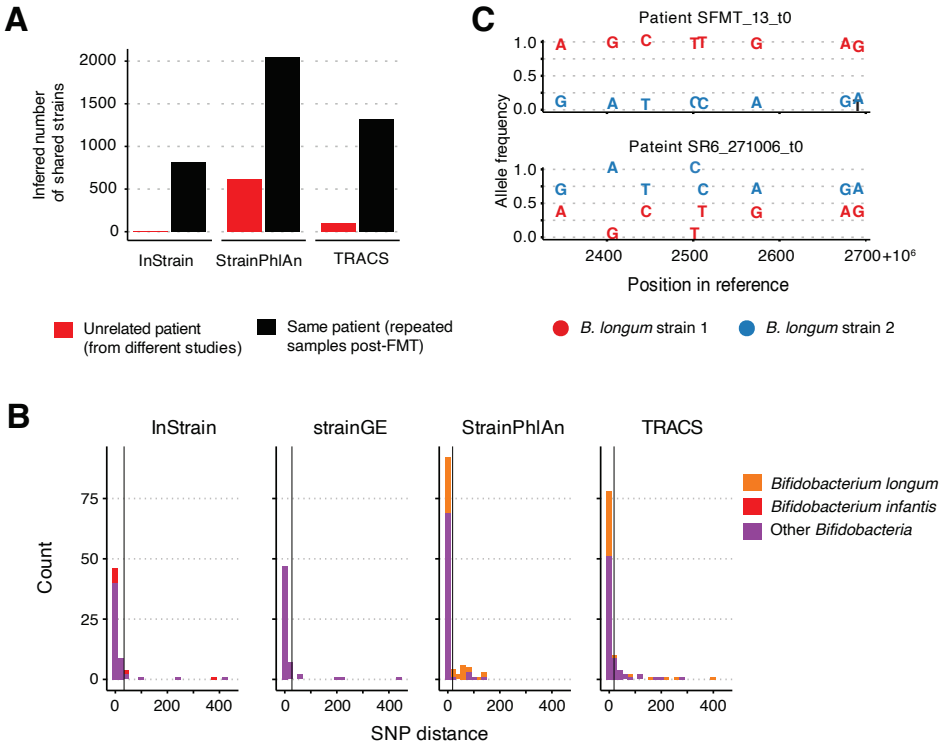



**Fig. 4. Estimates of engraftment in FMT triads.** A) The inferred number of plausible transmission pairs (shared strains), inferred by each algorithm, between samples taken as part of study investigating the impact of Faecal Microbiota Transplants (FMTs). Transmission between samples from unrelated participants (shown in red) are highly likely to be false positives. In contrast, samples from the same participant following FMT (shown in black) are more likely to be true positives, with an error rate proportional to that observed in the unrelated participant results. Species specific SNP thresholds for identifying recent transmission were inferred using the mixture distribution approach (see Methods). An analogous plot using the Youden method as described in Valles-Colomer et al.(2), is given in Supplementary Figure 1. B) Histograms indicating the distribution of pairwise SNP distances (truncated at 500 SNPs) between Donor and recipient (post-transplant) samples for major Bifidobacterial species. Vertical grey lines indicate the SNP thresholds inferred using the TRACS mixture distribution method. InStrain identified the transmission of *B. infantis*, which is likely *B. longum* but mislabeled in the UHGG genome collection database, whereas StrainGE identified no transmission of *B. longum*. C) An example of multiple strains of *B. longum* being transmitted between donor and recipient was detected exclusively by TRACS, whereas other algorithms failed to identify it. The allele frequencies at polymorphic loci within a segment of the *B. longum* reference genome are displayed for each sample. The inversion in strain frequency poses a challenge for alternative algorithms.

### Characterisation of maternal strain transmission and persistence across a large UK infant birth cohort.

The transmission and subsequent colonisation events that drive the development of our gut microbiota can have important implications on our health in childhood and later life. In particular, the impact of perturbations driven by interventions in early childhood, such as caesarean section and antimicrobial treatment, have been associated with health complications including asthma and atopy (35, 36).

To characterise the transmission and persistence of bacteria at the strain level during early childhood, we analysed faecal samples from 1,288 healthy, full-term infants, sequenced as part of the UK BabyBiome Study(32, 33). Faecal samples were collected from all babies at least once during the neonatal period ($\leq$ 1 month), including 29 twins and one triplet, with subsequent sampling from 302 infants, including 29 twins and one triplet, during later infancy (8.75 ± 1.98 months). Maternal faecal samples were also taken from 175 mothers, corresponding to 178 neonates. The TRACS algorithm was run on all samples using the GTDB reference database and SNP thresholds were calculated to distinguish recent transmission at the species level using the mixture distribution approach (as described in Methods). Consistent with the low error rate of TRACS and stringent infection control procedures in each hospital, the occurrence of shared strains between unrelated children, both within the same hospital and across different hospitals, was very low (0.78% and 0.73%, respectively; Figure 5A). Interestingly, the highest rate of strain sharing occurred between siblings, reflecting their closely shared strain colonisation history (from a common reservoir or between each other).

Infants born via caesarean section exhibited markedly reduced strain sharing compared to those delivered vaginally (Figure 5B), aligning with findings from previous studies (32, 33, 37). *Bifidobacterium bifidum*, *Bifidobacterium longum*, *Phocaeicola vulgatus*, *Parabacterioides disasonis* and *E. coli* showed some of the largest differences in transmission rates between mothers and infants born vaginally compared to those born via caesarean section. Consistent with previous results based on StrainPhlAn(32, 33, 37), maternal transmission of *Enterococcus faecalis* is extremely rare, with only one observed event between the mother and baby born via caesarean section. Previous studies have identified increased rates of colonisation with hospital associated species such as *E. faecalis*, *E. faecium* and *Klebsiella pneumoniae* in babies delivered by caesarean section (38). It is possible that the identified *E. faecalis* transmission resulted from separate colonisation events of the baby and the mother, originating

from the other unsampled sites such as the hospital environment. Overall, these results confirm the well-established patterns of maternal strain transmission observed with alternative tools, demonstrating the robustness of TRACS.

In addition to tracking strain transmission between hosts, TRACS can monitor the longitudinal persistence of strains within a single host. This includes observing the strain stability in infants, which was found to vary significantly between species (Figure 5C). Strains of commonly pathogenic species such as *E. coli* and *E. faecalis* were maintained for relatively short periods of time and are generally eliminated or replaced by late infancy. The persistence of the pioneer colonisers in the neonatal gut, including *B. longum* and *B. breve* strains(33), remained high during the first three weeks of infancy but declined markedly in late infancy, likely due to changes in diet and decreased breastfeeding rates (39, 40). In contrast, some other maternally transmitted commensal genera such as *B. bifidum* and closely related Phocaeicola, Bacteroides and Parabacteroides species were maintained at higher rates for the entire duration of the study. Critically, the TRACS algorithm identified 40.5% (49/121) additional strain sharing events compared to StrainPhlAn. This indicates instances of repeated co-colonisation of infants with multiple strains of *B. breve*, which were overlooked by algorithms like StrainPhlAn that consider only the dominant genotype (Supplementary Figure 3). For instance, Figure 5D and 5E illustrate the persistence of two *B. breve* strains over three time points within a single infant. Initially, the purple strain is dominant with a mean allele frequency of 0.82 (±0.081) but then drops to lower frequencies of 0.063 (±0.181) and 0.249 (±0.156) respectively at later time points. Taken together, these results demonstrate the effectiveness of the TRACS algorithm in analysing the fine-scale transmission and persistence of strains within the human microbiota.

## Discussion

TRACS is a versatile and modular algorithm for inferring recent pairwise transmission events from metagenomic and single species population sequencing data. Unlike alternative tools, TRACS was designed specifically for transmission inference. The algorithm addresses several key sources of error, including uneven sequencing coverage, shared sequence homology and horizontal gene transfer between microbes. Importantly, TRACS also accounts for the presence of multiple strains of the same species within a sample. Using simulations and artificial laboratory mixtures of *S. pneumoniae*, we show that accounting for shared homology and recombination substantially improves the accuracy of TRACS compared to alternative algorithms.

The scalability and versatility of the TRACS algorithm allows it to be applied to a broad spectrum of pathogens including viruses, bacteria and parasites. This includes contexts where species specific bioinformatics pipelines are yet to be developed. This was demonstrated with its application to SARS-CoV-2 data, where TRACS performed comparably to tailored approaches without requiring pre-specified filters for problematic sites. Although whole genome sequencing is not typically used to track the transmission of malaria-causing parasites, we demonstrate using single-cell genomics, that by omitting the complex and error-prone step of deconvolution, the TRACS algorithm effectively identifies shared strains of *P. falciparum* between samples. By estimating a lower bound on the SNP distance between strains in two samples, TRACS can reliably exclude recent transmission events. Combined with its versatility, computational efficiency, and ability to incorporate sampling dates through an enhanced version of the TransCluster model, TRACS is highly suitable for use in routine public health surveillance pipelines.

Beyond pathogen surveillance, TRACS can track the colonisation and carriage dynamics of the human microbiota. Using well-characterised metagenomic datasets from FMT studies, we demonstrate that the TRACS algorithm, which includes a novel method for estimating transmission distance thresholds, offers a superior balance of sensitivity and specificity for tracking strain acquisition and persistence within the gut. Notably, TRACS identifies instances of minority strain transmission between FMT donors and recipients, as well as longitudinally sampled infants, which are not detected by alternative tools. Moreover, in addition to confirming that caesarean sections significantly reduce mother-to-baby bacterial transmission, we identified species-specific differences in colonisation and persistence rates. Notably, *E. faecalis* and *B. cellulosilyticus* had higher transmission rates following caesarean delivery. As with other genomic transmission studies, it is essential to consider the broader epidemiology of host interactions. For example, in some cases, both mother and infant may be independently colonised by the same strain from the hospital environment or other unsampled sources, such as family members.

TRACS was developed to identify recent transmission events, enhancing both accuracy and ease of use. However, the algorithm was not developed for identifying high-quality within-host variants, making it unsuitable for detecting selection through dN/dS ratios. Furthermore, since the algorithm provides only a lower bound estimate of transmission or SNP distance, it cannot accurately infer genetic relationships over thousands of SNPs that represent much longer evolutionary timescales. For these purposes, tools like inStrain and MetaPhlAn respectively, are more appropriate. Similar to other metagenomic transmission inference methods, TRACS does not determine transmission direction. For this, we recommend using TRACS to identify potential transmission events, followed by more computationally intensive phylogenetic methods that incorporate additional epidemiological information (6, 41).

Metagenomics and deep population sequencing present a powerful alternative to relying solely on representative genomes for tracking the transmission and persistence of both pathogenic and commensal microbes. The versatility and robustness of the TRACS algorithm allows for the simultaneous analysis of multiple strains and species, which will help advance our understanding of how microbes transmit, establish, and persist within their hosts.

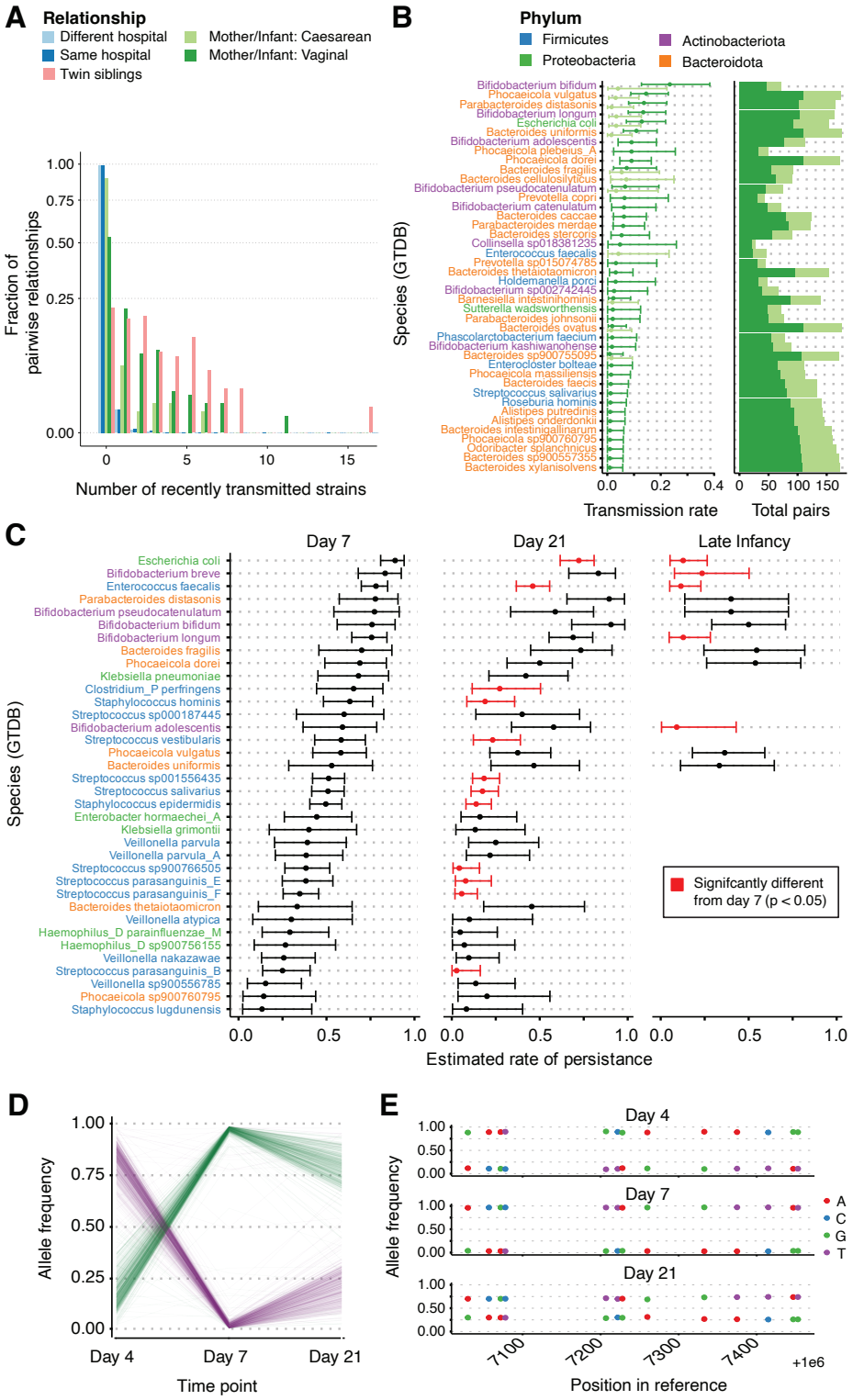



**Fig. 5. Maternal strain transmission and persistence across a large UK infant birth cohort.** A) A bar plot indicating the fraction of pairwise relationships that involve a putative recent transmission versus the number of shared strains. SNP distances and species-specific thresholds are inferred using the TRACS algorithm. Relationships between hosts are represented by different colours. The low false positive rate of transmission is evident, as the vast majority of inter-hospital relationships involve zero transmissions. B) Species-specific transmission rates between mothers and babies. Point estimates and error bars are inferred using a binomial model of transmission if a mother is colonised (see Methods). Points are coloured by the mode of delivery (as in A), highlighting the species-specific impact of delivery. The total number of possible transmission pairs (mothers colonised) is shown in the accompanying horizontal bar plot. C) The rate of strain persistence by species from birth to 7 days, 21 days, and late infancy. Persistence rates that differ significantly from day 7 are shown in red. Estimates are displayed only when at least 20 babies, who were originally observed to carry a species, are sampled again at respective time points. D) An example demonstrating the persistence of multiple distinct strains of *B. breve* in a single infant (ID: B00560) is presented. This example was not identified by StrainPhlAn, which only considers the dominant genotype. The frequencies of 1000 randomly chosen alleles, found to be at intermediate frequency on day 4, are shown. Initially, the purple strain dominates, but by days 7 and 21, the green strain becomes dominant. E) Plotting of individual allele frequencies of two *B. breve* strains (from panel D) relative to a portion of the *B. breve* reference genome.

## Methods

### TRACS algorithm.

The TRACS algorithm consists of three central modules that can be used independently (see Figure 1). The **alignment** module generates a polymorphism aware Multiple Sequence Alignment for each reference genome found within a given sample. Given an MSA, the **distance** module estimates pairwise SNP distances. Optionally, the distance module can also estimate the expected number of intermediate hosts separating pairs of samples using an extended version of the TransCluster algorithm(15). The TransCluster algorithm requires estimates of the transmission rate (transmissions/year) and clock rate (SNPs/genome/year) of the species to be provided. Finally, the **cluster** module estimates putative transmission clusters using single-linkage hierarchical clustering.

### Alignment.

The alignment module takes raw sequence data as input and produces alignments to multiple reference genomes in FASTA format. IUPAC ambiguity codes are used to represent polymorphic sites in each alignment. The alignment module supports a variety of database formats tailored to specific applications. Users can provide a single reference, utilise a Sourmash database that incorporates Genbank genome IDs, or opt for a custom TRACS database. The latter includes both reference genomes and a corresponding Sourmash database, which can be constructed using the TRACS build-db command.

If a database containing multiple references is provided, the algorithm will first select those represented in a sample using the Sourmash gather command (12). Minimap2 (42) is then used to align all reads within a sample to each reference independently and a pileup of alleles identified at each site in the reference is generated using htsbox (43).

To control for variable sequencing coverage, we developed an empirical Bayes method. The counts of minority variants at each site were modelled using a multinomial Dirichlet distribution. This allows for the estimation of the posterior frequency of variants across an alignment by considering the average frequency of minority variants at polymorphic sites as follows. Assume $n_{i1}$ is the read count of the most frequent allele at polymorphic site $i$ in an alignment with $n_{i2}$, $n_{i3}$ and $n_{i4}$ being the counts of the 2nd, 3rd and 4th most frequent allele respectively. Here we make the simplifying assumption that the mixture of alleles from each strain in a sample results in total allele frequencies that are observed in a consistent order.

We assume a Dirichlet prior on the joint frequency ($\mathbf{p_i} = (p_{i1}, p_{i2}, p_{i3}, p_{i4})$) of these counts with parameter vector $\alpha$ such that the probability density can be written as:

$$p(\mathbf{p}) \sim D(\alpha_1, \alpha_2, \alpha_3, \alpha_4) = \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \prod_k p_k^{\alpha_k - 1}$$

Assuming the resulting read counts are drawn from a multinomial distribution independently at each site the marginal likelihood is the Dirichlet-multinomial distribution given by:

$$p(D|\alpha) = \prod_i \left( \frac{\Gamma(\sum_k \alpha_k)}{\Gamma(\sum_k n_{ik} + \sum_k \alpha_k)} \prod_k \frac{\Gamma(n_{ik} + \alpha_k)}{\Gamma(\alpha_k)} \right)$$

where $D$ indicates the vector of read counts at each site $i$: $D = \{\mathbf{x}_1, ..., \mathbf{x}_N\}$ where $\mathbf{x}_i = \{n_{i1}, n_{i2}, n_{i3}, n_{i4}\}$

Point estimates of the $\alpha$ parameters were made by maximising the marginal likelihood using the fixed point iteration method (44). The maximum a posteriori estimate of the frequency of each allele under this prior at each site can then be estimated as:

$$f_{ij} = \frac{\alpha_j + n_{ij}}{\sum_k n_{ik} + \sum_k \alpha_k}$$

After estimating posterior frequencies at all variable sites in an alignment, TRACS applies a frequency cutoff which requires a coverage of at least 5 reads by default. Sites falling below this threshold are then excluded from further consideration in subsequent estimates of SNP distances. Furthermore, to control for major outliers in sequencing coverage, regions with coverage exceeding 1.5 times the interquartile range above or below the quartiles, as identified by Tukey's method, are excluded (45).

### Distance.

Pairwise SNP distances are estimated using a bitset-based algorithm to enhance both the speed and memory efficiency of the process, whilst incorporating IUPAC ambiguity codes. Each sequence in a Multiple Sequence Alignment is represented as a matrix of bits (0's and 1's), where rows indicate the four possible alleles and columns indicate the position in the sequence. Ambiguous bases are then represented by multiple 1's within a single column. A series of OR operations are used to rapidly calculate the sites which share common alleles between each pair of samples.

To account for the impacts of recombination and shared homology between strains, we include an optional additional step to identify regions within the pairwise comparison that have an elevated rate of SNPs. This uses a similar scan statistic to that used in popular recombination aware phylogenetic reconstruction algorithms (14).

Initially, the rate of SNPs within the alignment is calculated as $r = \frac{d}{L}$, where $d$ is the total number of SNPs and $L$ is the alignment length. For each pairwise comparison, a window size $w_l$ is chosen such that the expected number of SNPs within the window is equal to one: $E(n|p, w_l) = 1$. The maximum window size is set to 10kb. At each SNP location, $i$, the number of SNPs, $s_i$, within the window centred at $i$ is calculated. Assuming a binomial distribution, the probability of observing at least $s_i$ SNPs is given by:

$$P(S \geq s_i) = 1 - \sum_{k=0}^{s_i} \binom{w_l}{k} p^k (1-p)^{w_l - k}$$

A p-value threshold is then specified to exclude SNPs that are centred within a window that has a significantly high rate of SNPs ($p = 0.05$ by default). A Bonferroni correction is

made to control for multiple testing (46). For SNPs that fall within a distance $w_l/2$ of either edge of the alignment, $w_l$ is truncated.

### Modified TransCluster algorithm.

While SNP distances can be useful in distinguishing recent transmission events, they do not incorporate the sampling times of the hosts being considered. To incorporate the sampling times, TRACS includes an extended version of the TransCluster algorithm that both improves its speed and allows for the estimation of the expected number of intermediate hosts separating two samples.

Assuming SNPs accumulate at rate $\gamma$ and an average epidemiological generation time (the time between transmission events) $\beta$, the TransCluster algorithm can be used to estimate the probability of $k$ intermediate hosts occuring along the transmission chain between a pair of sampled hosts (15). Given a SNP distance of $N$ and a time separation of $\delta$ between the two samples, Stimson et al.(15) showed that the distribution of $k$ intermediate hosts could be written as:

$$P(k|h,\delta) = \frac{e^{-\delta(\lambda+\beta)}\lambda^{N+1}\beta^k}{k!F(N)} \int_{h=0}^{\infty} e^{-h(\lambda+\beta)}(h+\delta)^k \sum_{i=0}^{N} h^i \left(\frac{\delta^{N-i}}{i!(N-i)!}\right) dh$$

where $h$ is the evolutionary time separating two samples. In the original implementation of TransCluster, this expression was computed via numerical integration. This method can become computationally prohibitive when considering thousands of pairwise comparisons. We show that this equation can be re-written as a finite sum that can be solved rapidly using efficient caching of intermediate values:

$$P(k|h,\delta) = \frac{\lambda^{N+1}\beta^k(N+k)!}{e^{\delta\beta}N!k!\sum_{i=0}^{N}\frac{(\lambda\delta)^i}{i!}} \sum_{i=0}^{N+k} \frac{\delta^{N+k-i}}{(N+k-i)!(\lambda+\beta)^{i+1}}$$

A full derivation is given in the supplementary methods. Given this derivation, the expected number of intermediate hosts can be written as:

$$\begin{aligned} E(K|h,\delta) &= \sum_{k=0}^{\infty} kP(k|h,\delta) \\ &= \sum_{k=0}^{\infty} k\frac{\lambda^{N+1}\beta^k(N+k)!}{e^{\delta\beta}N!k!\sum_{i=0}^{N}\frac{(\lambda\delta)^i}{i!}} \sum_{i=0}^{N+k} \frac{\delta^{N+k-i}}{(N+k-i)!(\lambda+\beta)^{i+1}} \end{aligned}$$

While this is an infinite sum, an upper bound on the error after truncating at a given $k$ can be calculated (supplementary methods) allowing for the sum to be calculated to a user specified precision. Again, TRACS makes use of efficient caching strategies to reduce the number of times this equation is calculated.

### Clustering.

Consistent with other transmission inference pipelines, TRACS implements single linkage hierarchical clustering to identify putative transmission clusters. The **cluster** module takes a list of pairwise distance estimates output from the **distance** module and a user supplied SNP or transmission distance threshold and outputs a CSV file indicating which genome belongs to each cluster. Guidance on the selection of an appropriate threshold is given below. Single linkage clustering is preferable as it allows for long transmission chains to be grouped into a single cluster.

### Inferring a transmission threshold.

To determine an appropriate SNP threshold for distinguishing recent transmission, we assume that it is possible to provide two sets of samples where the first includes possible transmission events and the second is unlikely to include any recent transmission. By fitting a mixture distribution to these samples, we can calculate a suitable threshold.

We assume that SNP distances between pairs of samples that are unlikely to be related by recent transmission ($d_u$) are distributed according to a Negative Binomial distribution with parameters $n$ and $p$ such that

$$p(d_u = x) = \frac{\Gamma(x+n)}{\Gamma(n)x!}p^n(1-p)^x$$

Moreover, we assume that the distribution of SNP distances between pairs of SNPs related by recent transmission ($d_r$) follows a Poisson distribution with mean $\lambda$

$$p(d_r = x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

Thus the distribution of SNP distances in the dataset that contains both recent transmission and more distant transmission ($d_m$) can be modelled as a mixture of these two distributions where $q$ indicates the probability that any given pair of samples are related by recent transmission.

$$p(d_m = x) = q * p(d_r = x) + (1-q) * p(d_u = x)$$

To estimate these parameters we fit the Negative Binomial model to the dataset containing exclusively distantly related samples using maximum likelihood. This produces a distribution of the SNP distance between distantly related samples based on the second dataset. Substituting these estimates into the mixture then allows for the estimation of the remaining parameters using maximum likelihood. Note that during this second estimation, we use the first data set without the second dataset. Finally a SNP threshold can be chosen by selecting the probability that a close transmission pair is erroneously excluded.

To select thresholds for the analysis of transmission in the mother and baby dataset(32, 33), repeated samples of the same baby were used as the 'close transmission' dataset and pairs of samples from different infants were used as the 'distant' dataset. In the FMT dataset, the 'close transmission' dataset consisted of samples from the same patient post-FMT and the 'distant' dataset consisted of pairs of samples taken from unrelated donors. In both cases, a probability of missing a transmission of 0.001 was chosen. To adjust for the fact that the 'close transmission' pairs were from the same patient, the estimated thresholds were multiplied by 3 to allow for twice the evolutionary time to occur in the mother/donor.

### Alternative algorithms.

As alternative algorithms do not always allow for the specification of custom databases, it can be challenging to compare the results of each approach.

On the simulated metagenomic data and FMT metagenomic data, the GTDB r207 bacterial reference database was used (34) to run TRACS. To consider both the simulated and real pneumococcal deep population sequencing, TRACS was run using representative genomes from each sequencing cluster described in the Global Pneumococcal Sequencing project (25). To analyse the SARS-CoV-2 and *P. falciparum* deep population sequencing data, the reference genome from each species was used when running TRACS.

In all cases, StrainPhlAn was run using version 4.0.5 with the default database using the command line parameters outlined in Valles-Colomer et al(2).

Since StrainGE is designed to run on species-specific databases, we used the species reference genomes from RefSeq for each species included in the metagenomics simulation data. As it was computationally expensive to consider all possible species when running StrainGE on the FMT data, we restricted all analyses to the Bifidobacterium genus. Reference genomes were downloaded from Refseq and the database was constructed following the StrainGE user manual.

On the simulated and FMT metagenomic data, InStrain was run using the approach outlined in the InStrain user manual. The 19F serotype included in the laboratory mixtures was used to run InStrain on the simulated pneumococcal deep population sequencing data.

To compare the resource requirements of each algorithm, we measured CPU and memory usage for a pair of simulated metagenomic samples where one genome was simulated to have been transmitted (Supplementary Figure 4). Comparing the results across algorithms is challenging, as each requires different tasks to be performed separately by the user, such as read alignment and SNP distance calculation. To ensure a fair comparison, we used the same tools employed by TRACS for operations not automatically handled by other pipelines. Although the exact differences in speed and memory usage may vary depending on the dataset, TRACS consistently demonstrated the lowest memory and CPU usage in this test, outperforming the next best algorithm by approximately a factor of 2 (Supplementary Figure 4).

The exact commands used to run StrainPhlAn, inStrain and StrainGE are given in the supplementary code provided in the associated GitHub repository (`https://github.com/gtonkinhill/tracs_manuscript`).

### Simulated datasets.

Given a user-specified SNP distance and a set of reference genomes, transmission of an individual strain between pairs of metagenomic or population sequencing samples were simulated using a custom python script provided as part of the TRACS package.

The number of reference genomes included in each sample was simulated using a Poisson distribution. The user specified number of SNPs was then simulated by randomly introducing mutations along one reference genome. This genome was included in both samples in a pair. The remaining reference genomes in each sample were randomly selected from the user provided list. In cases where both samples shared a reference genome other than the transmitted one, additional SNPs were randomly introduced. The total number of SNPs separating non-transmitted strains was drawn from a Poisson distribution with mean 10,000.

Simulated sequencing depths were drawn from a Dirichlet distribution with all parameters equal to 1. A cumulative read depth across all genomes of 500 reads was used and Illumina reads were generated using the ART synthetic read simulator (47). Simulated Nanopore reads were generated using Badread v0.4.1 (48). A full set of parameters and commands can be found in the supplementary code provided in the GitHub repository.

To simulate transmission between gut metagenomes, 11 bacterial reference genomes were selected from a study investigating the transmission of drug-resistant pathogens using metagenomics(20). To simulate transmission among deeply sequenced populations of pneumococci, we randomly selected 28 pneumococcal genomes from a prior study (49). A full list of all commands used along with the reference genomes is provided in the associated GitHub repository.

### Data availability.

SARS-CoV-2 sequencing reads were taken from a recent study investigating the within-host evolution of the virus(24). The ENA IDs of the 37 SARS-CoV-2 samples found to contain multiple distinct lineages are listed in the supplementary GitHub repository. All sequencing reads from the original study are available in the ENA under accession number ERP126512. The *S. pneumoniae* and *P. falciparum* sequencing reads are available from the ENA under project codes PRJEB22771 and PRJNA482776 respectively(5, 30). The deep metagenomic sequencing reads of gut samples from mothers and their children is available from the ENA under the accession number ERP115334(32, 33). The FMT triad samples included sequencing data published in the ENA under project code PRJEB47909(31). The sample accession codes used in all analyses are provided in the supplementary files on GitHub.

### Code availability.

The TRACS code and user documentation is available under an MIT open source licence on GitHub at `https://github.com/gtonkinhill/tracs/`. Supplementary code used to generate the analysis and figures in this manuscript is available at `https://github.com/gtonkinhill/tracs_manuscript`.

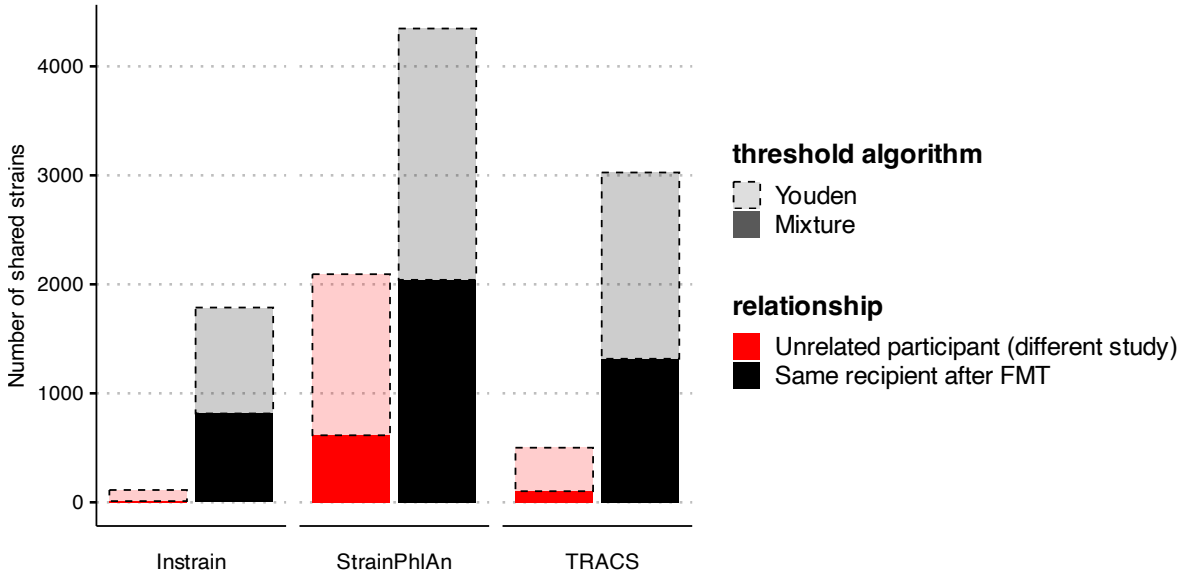
## References


1. Simon R Harris, Edward J Feil, Matthew T G Holden, Michael A Quail, Emma K Nickerson, Narisara Chantratita, Susana Gardete, Ana Tavares, Nick Day, Jodi A Lindsay, Jonathan D Edgeworth, Hermínia de Lencastre, Julian Parkhill, Sharon J Peacock, and Stephen D Bentley. Evolution of MRSA during hospital transmission and intercontinental spread. Science, 327(5964):469–474, January 2010.

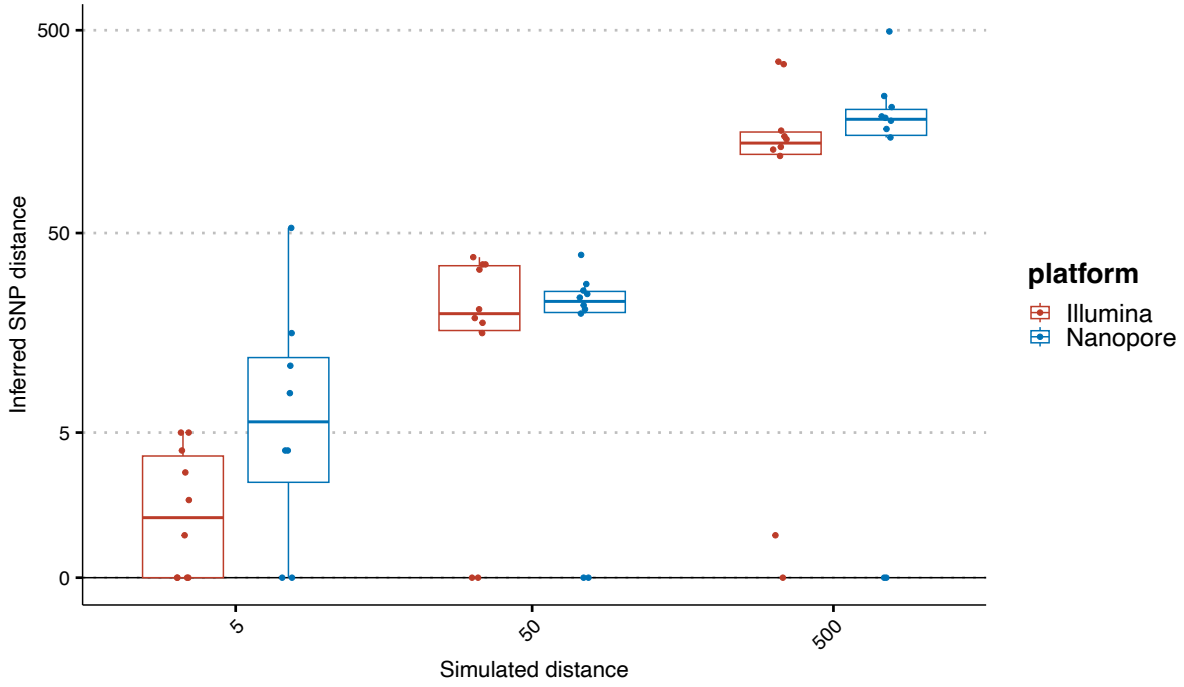
2. Mireia Valles-Colomer, Aitor Blanco-Míguez, Paolo Manghi, Francesco Asnicar, Leonard Dubois, Davide Golzato, Federica Armanini, Fabio Cumbo, Kun D Huang, Serena Manara, Giulia Masetti, Federica Pinto, Elisa Piperni, Michal Punčochář, Liviana Ricci, Moreno Zolfo, Olivia Farrant, Adriana Goncalves, Marta Selma-Royo, Ana G Binetti, Jimmy E Becerra, Bei Han, John Lusingu, John Amuasi, Loredana Amoroso, Alessia Visconti, Claire M Steves, Mario Falchi, Michele Filosi, Adrian Tett, Anna Last, Qian Xu, Nan Qin, Huanlong Qin, Jürgen May, Daniel Eibach, Maria Valeria Corrias, Mirco Ponzoni, Edoardo Pasolli, Tim D Spector, Enrico Domenici, Maria Carmen Collado, and Nicola Segata. The person-to-person transmission landscape of the gut and oral microbiomes. Nature, 614(7946): 125–135, February 2023.
3. Matthew R Olm, Alexander Crits-Christoph, Keith Bouma-Gregson, Brian A Firek, Michael J Morowitz, and Jillian F Banfield. instrain profiles population microdiversity from metagenomic data and sensitively detects shared microbial strains. Nat. Biotechnol., 39(6):727–736, June 2021.
4. Tami D Lieberman, Kelly B Flett, Idan Yelin, Thomas R Martin, Alexander J McAdam, Gregory P Priebe, and Roy Kishony. Genetic variation of a bacterial pathogen within individuals with cystic fibrosis provides a record of selective pressures. Nat. Genet., 46(1):82–87, January 2014.
5. Gerry Tonkin-Hill, Clare Ling, Chrispin Chaguza, Susannah J Salter, Pattaraporn Hinfonthong, Elissavet Nikolaou, Natalie Tate, Andrzej Pastusiak, Claudia Turner, Claire Chewapreecha, Simon D W Frost, Jukka Corander, Nicholas J Croucher, Paul Turner, and Stephen D Bentley. Pneumococcal within-host diversity during colonization, transmission and treatment. Nat Microbiol, 7(11):1791–1804, November 2022.
6. Chris Wymant, Matthew Hall, Oliver Ratmann, David Bonsall, Tanya Golubchik, Mariateresa de Cesare, Astrid Gall, Marion Cornelissen, Christophe Fraser, and STOP-HCV Consortium, The Maela Pneumococcal Collaboration, and The BEEHIVE Collaboration. PHYLOSCANNER: Inferring transmission from within- and Between-Host pathogen genetic diversity. Mol. Biol. Evol., November 2017.
7. Duy Tin Truong, Adrian Tett, Edoardo Pasolli, Curtis Huttenhower, and Nicola Segata. Microbial strain-level population structure and genetic diversity from metagenomes. Genome Res., 27(4):626–638, April 2017.
8. Lucas R van Dijk, Bruce J Walker, Timothy J Straub, Colin J Worby, Alexandra Grote, Henry L Schreiber, 4th, Christine Anyansi, Amy J Pickering, Scott J Hultgren, Abigail L Manson, Thomas Abeel, and Ashlee M Earl. StrainGE: a toolkit to track and characterize low-abundance strains in complex microbial communities. Genome Biol., 23(1):74, March 2022.
9. Chunyu Zhao, Boris Dimitrov, Miriam Goldman, Stephen Nayfach, and Katherine S Pollard. MIDAS2: Metagenomic intra-species diversity analysis system. Bioinformatics, 39(1), January 2023.
10. Stephen Nayfach, Beltran Rodriguez-Mueller, Nandita Garud, and Katherine S Pollard. An integrated metagenomics pipeline for strain profiling reveals novel patterns of bacterial transmission and biogeography. Genome Res., 26(11):1612–1625, November 2016.
11. Christopher Quince, Sergey Nurk, Sebastien Raguideau, Robert James, Orkun S Soyer, J Kimberly Summers, Antoine Limasset, A Murat Eren, Rayan Chikhi, and Aaron E Darling. STRONG: metagenomics strain resolution on assembly graphs. Genome Biol., 22(1):214, July 2021.
12. N Tessa Pierce, Luiz Irber, Taylor Reiter, Phillip Brooks, and C Titus Brown. Large-scale sequence comparisons with sourmash. F1000Res., 8:1006, July 2019.
13. Heng Li. Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. Bioinformatics, 32(14):2103–2110, July 2016.
14. Nicholas J Croucher, Andrew J Page, Thomas R Connor, Aidan J Delaney, Jacqueline A Keane, Stephen D Bentley, Julian Parkhill, and Simon R Harris. Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using gubbins. Nucleic Acids Res., page gku1196, November 2014.
15. James Stimson, Jennifer Gardy, Barun Mathema, Valeriu Crudu, Ted Cohen, and Caroline Colijn. Beyond the SNP threshold: Identifying outbreak clusters using inferred transmissions. Mol. Biol. Evol., 36(3):587–603, March 2019.
16. Torsten Seemann. snippy.
17. W J Youden. Index for rating diagnostic tests. Cancer, 3(1):32–35, January 1950.
18. Finlay Campbell, Camilla Strang, Neil Ferguson, Anne Cori, and Thibaut Jombart. When are pathogen genome sequences informative of transmission events? PLoS Pathog., 14 (2):e1006885, February 2018.
19. James R Knight, Eileen M Dunne, E Kim Mulholland, Sudipta Saha, Catherine Satzke, Adrienn Tothpal, and Daniel M Weinberger. Determining the serotype composition of mixed samples of pneumococcus using whole-genome sequencing. Microb Genom, 7(1), January 2021.
20. Andre Mu, Jason C Kwong, Nicole S Isles, Anders Gonçalves da Silva, Mark B Schultz, Susan A Ballard, Courtney R Lane, Glen P Carter, Deborah A Williamson, Torsten Seemann, Timothy P Stinear, and Benjamin P Howden. Reconstruction of the genomes of Drug-Resistant pathogens for outbreak investigation through metagenomic sequencing. mSphere, 4(1), January 2019.
21. Katrina A Lythgoe, Matthew Hall, Luca Ferretti, Mariateresa de Cesare, George MacIntyre-Cockett, Amy Trebes, Monique Andersson, Newton Otecko, Emma L Wise, Nathan Moore, Jessica Lynch, Stephen Kidd, Nicholas Cortes, Matilde Mori, Rebecca Williams, Gabrielle Vernet, Anita Justice, Angie Green, Samuel M Nicholls, M Azim Ansari, Lucie Abeler-Dörner, Catrin E Moore, Timothy E A Peto, David W Eyre, Robert Shaw, Peter Simmonds, David Buck, John A Todd, Oxford Virus Sequencing Analysis Group (OVSG), Thomas R Connor, Shirin Ashraf, Ana da Silva Filipe, James Shepherd, Emma C Thomson, COVID-19 Genomics UK (COG-UK) Consortium, David Bonsall, Christophe Fraser, and Tanya Golubchik. SARS-CoV-2 within-host diversity and transmission. Science, 372(6539), April 2021.
22. Nicola De Maio, Conor Walker, Rui Borges, Lukas Weilguny, Greg Slodkowicz, and Nick Goldman. Issues with SARS-CoV-2 sequencing data. https://virological.org/t/issues-with-sars-cov-2-sequencing-data/473, May 2020. Accessed: 2024-5-8.
23. D N A Pipelines, Benjamin Farr, Diana Rajan, Emma Betteridge, Lesley Shirley, Michael Quail, Naomi Park, Nicholas Redshaw, Iraad Bronner, Louise Aigrain, Scott Goodwin, Scott Thurston, Stefanie Lensing, James Bonfield, Keith James, Nicholas Salmon, Charlotte Beaver, Rachel Nelson, K David, Alex Alderton, and Ian Johnston. COVID-19 ARTIC v3 illumina library construction and sequencing protocol v4. Technical report, May 2020.
24. Gerry Tonkin-Hill, Inigo Martincorena, Roberto Amato, Andrew R J Lawson, Moritz Gerstung, Ian Johnston, David K Jackson, Naomi Park, Stefanie V Lensing, Michael A Quail, Sónia Gonçalves, Cristina Ariani, Michael Spencer Chapman, William L Hamilton, Luke W Meredith, Grant Hall, Aminu S Jahun, Yasmin Chaudhry, Myra Hosmillo, Malte L Pinckert, Iliana Georgana, Anna Yakovleva, Laura G Caller, Sarah L Caddy, Theresa Feltwell, Fahad A Khokhar, Charlotte J Houldcroft, Martin D Curran, Surendra Parmar, COVID-19 Genomics UK (COG-UK) Consortium, Alex Alderton, Rachel Nelson, Ewan M Harrison, John Sillitoe, Stephen D Bentley, Jeffrey C Barrett, M Estee Torok, Ian G Goodfellow, Cordelia Langford, Dominic Kwiatkowski, and Wellcome Sanger Institute COVID-19 Surveillance Team. Patterns of within-host genetic diversity in SARS-CoV-2. Elife, 10:e66857, August 2021.
25. Rebecca A Gladstone, Stephanie W Lo, John A Lees, Nicholas J Croucher, Andries J van Tonder, Jukka Corander, Andrew J Page, Pekka Marttinen, Leon J Bentley, Theresa J Ochoa, Pak Leung Ho, Mignon du Plessis, Jennifer E Cornick, Brenda Kwambana-Adams, Rachel Benisty, Susan A Nzenze, Shabir A Madhi, Paulina A Hawkins, Dean B Everett, Martin Antonio, Ron Dagan, Keith P Klugman, Anne von Gottberg, Lesley McGee, Robert F Breiman, and Stephen D Bentley. International genomic definition of pneumococcal lineages, to contextualise disease, antibiotic resistance and vaccine impact. EBioMedicine, 43:338–346, May 2019.
26. John A Lees, Nicholas J Croucher, David Goldblatt, François Nosten, Julian Parkhill, Claudia Turner, Paul Turner, and Stephen D Bentley. Genome-wide identification of lineage and locus specific variation associated with pneumococcal carriage duration. Elife, 6, July 2017.
27. Sophie Belman, Henri Pesonen, Nicholas J Croucher, Stephen D Bentley, and Jukka Corander. Estimating between country migration in pneumococcal populations. G3, March 2024.
28. World malaria report 2023. Technical report, World Health Organization, 2023.
29. Sha Joe Zhu, Jason A Hendry, Jacob Almagro-Garcia, Richard D Pearson, Roberto Amato, Alistair Miles, Daniel J Weiss, Tim Cd Lucas, Michele Nguyen, Peter W Gething, Dominic Kwiatkowski, Gil McVean, and Pf3k Project. The origins and relatedness structure of mixed infections vary with local prevalence of p. falciparum malaria. Elife, 8, July 2019.
30. Standwell C Nkhoma, Simon G Trevino, Karla M Gorena, Shalini Nair, Stanley Khoswe, Catherine Jett, Roy Garcia, Benjamin Daniel, Aliou Dia, Dianne J Terlouw, Stephen A Ward, Timothy J C Anderson, and Ian H Cheeseman. Co-transmission of related malaria parasite lineages shapes Within-Host parasite diversity. Cell Host Microbe, 27(1):93–103.e4, January 2020.
31. Gianluca Ianiro, Michal Punčochář, Nicolai Karcher, Serena Porcari, Federica Armanini, Francesco Asnicar, Francesco Beghini, Aitor Blanco-Míguez, Fabio Cumbo, Paolo Manghi, Federica Pinto, Luca Masucci, Gianluca Quaranta, Silvia De Giorgi, Giusi Desirè Sciumè, Stefano Bibbò, Federica Del Chierico, Lorenza Putignani, Maurizio Sanguinetti, Antonio Gasbarrini, Mireia Valles-Colomer, Giovanni Cammarota, and Nicola Segata. Variability of strain engraftment and predictability of microbiome composition after fecal microbiota transplantation across different diseases. Nat. Med., 28(9):1913–1923, September 2022.
32. Yan Shao, Samuel C Forster, Evdokia Tsaliki, Kevin Vervier, Angela Strang, Nandi Simpson, Nitin Kumar, Mark D Stares, Alison Rodger, Peter Brocklehurst, Nigel Field, and Trevor D Lawley. Stunted microbiota and opportunistic pathogen colonization in caesarean-section birth. Nature, 574(7776):117–121, September 2019.
33. Yan Shao, Cristina Garcia-Mauriño, Simon Clare, Nicholas J R Dawson, Andre Mu, Anne Adoum, Katherine Harcourt, Junyan Liu, Hilary P Browne, Mark D Stares, Alison Rodger, Peter Brocklehurst, Nigel Field, and Trevor D Lawley. Primary succession of bifidobacteria drives pathogen resistance in neonatal microbiota assembly. Nat. Microbiol., In press, 2024.
34. Donovan H Parks, Maria Chuvochina, Christian Rinke, Aaron J Mussig, Pierre-Alain Chaumeil, and Philip Hugenholtz. GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. Nucleic Acids Res., 50(D1):D785–D794, January 2022.
35. Kei E Fujimura, Alexandra R Sitarik, Suzanne Havstad, Din L Lin, Sophia Levan, Douglas Fadrosh, Ariane R Panzer, Brandon LaMere, Elze Rackaityte, Nicholas W Lukacs, Ganesa Wegienka, Homer A Boushey, Dennis R Ownby, Edward M Zoratti, Albert M Levin, Christine C Johnson, and Susan V Lynch. Neonatal gut microbiota associates with childhood multisensitized atopy and T cell differentiation. Nat. Med., 22(10):1187–1191, October 2016.
36. Jakob Stokholm, Martin J Blaser, Jonathan Thorsen, Morten A Rasmussen, Johannes Waage, Rebecca K Vinding, Ann-Marie M Schoos, Asja Kunøe, Nadia R Fink, Bo L Chawes, Klaus Bønnelykke, Asker D Brejnrod, Martin S Mortensen, Waleed Abu Al-Soud, Søren J Sørensen, and Hans Bisgaard. Maturation of the gut microbiome and risk of asthma in childhood. Nat. Commun., 9(1):141, January 2018.
37. Léonard Dubois, Mireia Valles-Colomer, Alise Ponsero, Otto Helve, Sture Andersson, Kaija-Leena Kolho, Francesco Asnicar, Katri Korpela, Anne Salonen, Nicola Segata, and Willem M de Vos. Paternal and induced gut microbiota seeding complement mother-to-infant transmission. Cell Host Microbe, 32(6):1011–1024.e4, June 2024.
38. Hakdong Shin, Zhiheng Pei, Keith A Martinez, 2nd, Juana I Rivera-Vinas, Keimari Mendez, Humberto Cavallin, and Maria G Dominguez-Bello. The first microbial environment of infants born by c-section: the operating room microbes. Microbiome, 3:59, December 2015.
39. Dena Ennis, Shimrit Shmorak, Evelyn Jantscher-Krenn, and Moran Yassour. Longitudinal quantification of bifidobacterium longum subsp. infantis reveals late colonization in the infant gut independent of maternal milk HMO composition. Nat. Commun., 15(1):894, January 2024.
40. Marta Selma-Royo, Léonard Dubois, Serena Manara, Federica Armanini, Raúl Cabrera-Rubio, Mireia Valles-Colomer, Sonia González, Anna Parra-Llorca, Ramon Escuriet, Lars Bode, Cecilia Martínez-Costa, Nicola Segata, and Maria Carmen Collado. Birthmode and environment-dependent microbiota transmission dynamics are complemented by breastfeeding during the first year. Cell Host Microbe, 32(6):996–1010.e4, June 2024.
41. Nicola De Maio, Colin J Worby, Daniel J Wilson, and Nicole Stoesser. Bayesian reconstruction of transmission within outbreaks using genomic variants. PLoS Comput. Biol., 14(4): e1006117, April 2018.
42. Heng Li. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics, 34(18):

3094–3100, September 2018.
43. Heng Li. htsbox, 2013.
44. Thomas P Minka. Estimating a dirichlet distribution. http://citeseerx.ist.psu.edu › viewdoc › summaryhttp://citeseerx.ist.psu.edu › viewdoc › summary, 2000.
45. John Wilder Tukey. Exploratory Data Analysis. Addison-Wesley Publishing Company, 1977.
46. Carlo Bonferroni. Teoria statistica delle classi e calcolo delle probabilita. Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commericiali di Firenze, 8:3–62, 1936.
47. Weichun Huang, Leping Li, Jason R Myers, and Gabor T Marth. ART: a next-generation sequencing read simulator. Bioinformatics, 28(4):593–594, February 2012.
48. Ryan Wick. Badread: simulation of error-prone long reads. J. Open Source Softw., 4(36): 1316, April 2019.
49. Claire Chewapreecha, Simon R Harris, Nicholas J Croucher, Claudia Turner, Pekka Marttinen, Lu Cheng, Alberto Pessia, David M Aanensen, Alison E Mather, Andrew J Page, Susannah J Salter, David Harris, Francois Nosten, David Goldblatt, Jukka Corander, Julian Parkhill, Paul Turner, and Stephen D Bentley. Dense genomic sampling identifies highways of pneumococcal recombination. Nat. Genet., 46(3):305–309, March 2014.

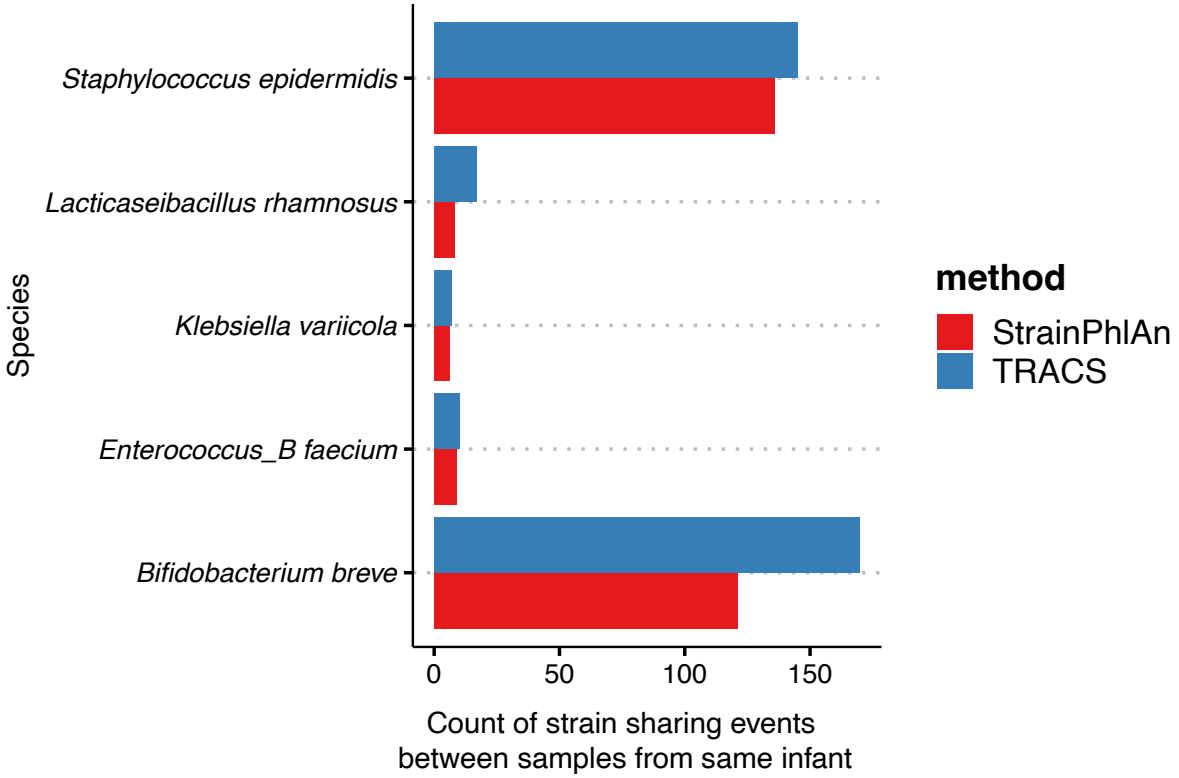
## Supplementary Figures



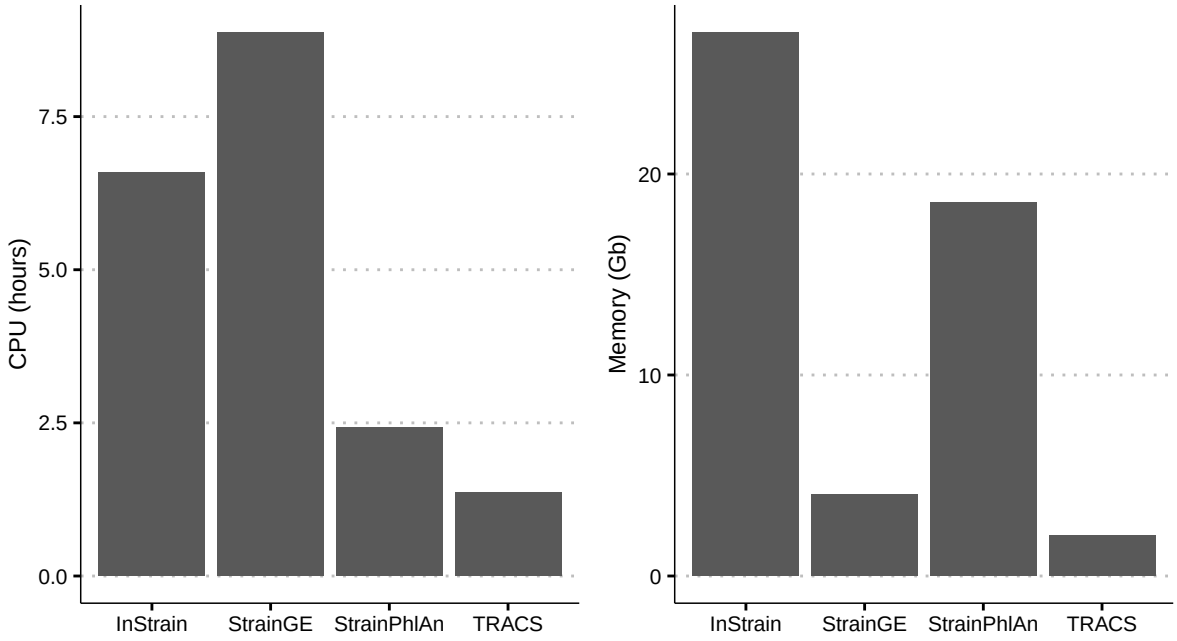

**Supplementary Figure 1.** The inferred number of transmission pairs, inferred by each algorithm, between samples taken as part of a study investigating the impact of Faecal Microbiota Transplants (FMTs)(31). Transmission between samples from unrelated participants (shown in red) are highly likely to be false positives. In contrast, samples from the same participant following FMT (shown in black) are more likely to be true positives, with an error rate proportional to that observed in the unrelated participant results. Species specific SNP thresholds for identifying recent transmission were inferred using the mixture distribution approach (solid colour). Additional transmission pairs identified using the Youden cut-off method proposed in Valles-Collomer et al.(2), are shown in a lighter colour with a dotted outline. This indicates that the Youden approach is substantially more relaxed and is likely to result in a high rate of false positives.




**Supplementary Figure 2.** The inferred versus simulated distance of running TRACS on simulated mixtures of pneumococcal strains, where each sample pair is simulated to have a single strain transmitted at the given SNP distance. Here, both Illumina (red) and Nanopore R10.4.1 (blue) sequencing reads have been simulated using the ART and Badread simulators, respectively. Boxplots indicate the distribution of the inferred number of SNPS. The central line represents the median, the edges of the box indicate the interquartile range (IQR), and the whiskers extend to 1.5 times the IQR. This indicates that TRACS can be used with high accuracy long read sequencing technology.




**Supplementary Figure 3.** Species found to have higher rates of persistence when analysed using TRACS (blue) compared to the StrainPhlAn (red) algorithm as originally used in the initial analysis of this cohort (32). Due to the substantially high rate of false positives associated with StrainPhlan it is likely that additional species may experience frequent colonisation by multiple strains which are not shown here.




**Supplementary Figure 4.** The CPU and memory requirements for running each pipeline on a simulated pair of metagenomic samples where a single genome was simulated to have transmitted. These results offer a general indication of each algorithm's performance. However, the exact differences in speed and memory usage will vary depending on the characteristics of each specific dataset.