Metagenomics Strain Resolution on Assembly Graphs

Christopher Quince[1,*], Sergey Nurk[2,*], Sebastien Raguideau[1], Robert James[3], Orkun S. Soyer[3], J. Kimberly Summers[1], Antoine Limasset[4], A. Murat Eren[5,6], Rayan Chikhi[7], Aaron E. Darling[8]

1 Warwick Medical School, University of Warwick, Gibbet Hill Road, Coventry, CV4 7AL, UK
2 Genome Informatics Section, Computational and Statistical Genomics Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, 20892, USA
3 School of Life Sciences, University of Warwick, Gibbet Hill Road, Coventry, CV4 7AL, UK
4 Univ. Lille, CNRS, Inria, UMR 9189 - CRIStAL, France
5 Department of Medicine, University of Chicago, Chicago, Illinois, USA
6 Josephine Bay Paul Center, Marine Biological Laboratory, Woods Hole, Massachusetts, USA
7 Department of Computational Biology, Institut Pasteur, C3BI USR 3756 IP CNRS, Paris, France
8 The ithree institute, University of Technology Sydney, 15 Broadway, Ultimo, 2007, NSW, Australia

* Joint corresponding authors: c.quince@warwick.ac.uk and sergey.nurk@nih.gov

Abstract

We introduce a novel bioinformatics pipeline, STrain Resolution ON assembly Graphs (STRONG), which identifies strains *de novo*, when multiple metagenome samples from the same community are available. STRONG performs coassembly, followed by binning into metagenome assembled genomes (MAGs), but uniquely it stores the coassembly graph prior to simplification of variants. This enables the subgraphs for individual single-copy core genes (SCGs) in each MAG to be extracted. It can then thread back reads from the samples to compute per sample coverages for the unitigs in these graphs. These graphs and their unitig coverages are then used in a Bayesian algorithm, BayesPaths, that determines the number of strains present, their sequences or haplotypes on the SCGs and their abundances in each of the samples.

Our approach both avoids the ambiguities of read mapping and allows more of the information on co-occurrence of variants in reads to be utilised than if variants were treated independently, whilst at the same time exploiting the correlation of variants across samples that occurs when they are linked in the same strain. We compare STRONG to the current state of the art on synthetic communities and demonstrate that we can recover more strains, more accurately, and with a realistic estimate of uncertainty deriving from the variational Bayesian algorithm employed for the strain resolution. On a real anaerobic digestor time series we obtained strain-resolved SCGs for over 300 MAGs that for abundant community members match those observed from long Nanopore reads.

Keywords: microbiome, metagenome, strains, Bayesian, microbial community, assembly graph

Introduction

There is a growing realisation that to fully understand microbial communities it is necessary to resolve them to the level of individual strains [35]. The strain is for many species the fundamental unit of microbiological diversity. This is because two strains of the same species can have very different functional roles. The classic example is *E. coli*, where one strain can be a dangerous pathogen and another a harmless commensal [24]. The best definition of a strain, and the only one that avoids ambiguity, is a set of clonal descendants of a single cell [15, 39], but strain genomes by this definition can only reliably be determined by sequencing cultured isolates or single cells [30]. The former is not representative of the community and the latter is still too expensive and low-throughput for many applications as well as producing only fragmentary genomes. For these reasons, there is a practical need for efficient methods that can profile microbial communities at high genomic resolution.

In contrast to 16S rRNA gene sequencing, shotgun metagenomics has the potential to resolve microbial communities to the strain level. This is because it generates reads from throughout the genomes of all the community members. It also has the additional advantages of reduced levels of bias and the capability to reconstruct genomes. There are many methods for reference-based strain resolution from metagenome data [1, 35, 42], but they are, and will continue to be, limited by the challenge of comprehensively isolating and sequencing the genomes of diverse microbial strains. Comprehensive reference genome databases may be possible for a few slowly evolving species or particularly well studied pathogens but for the entirety of a complex community it is unlikely to ever be tractable. For example, in a recent *de novo* large-scale binning study of the relatively well-studied human gut microbiome, it was found that 77% of the species recovered did not have a reference genome in public databases [31]. This suggests that even less of the strain-level diversity in those samples would be represented in a genome database. These observations motivate the need for *de novo* methods of metagenomic strain resolution.

In the metagenomics context, we adopt the definition of a 'metagenome strain' as a clonal subpopulation with sufficiently low levels of recombination with other strains, that it can be distinguished genetically from them. This does not require that recombination between strains does not occur, rather that either because of physical separation or selection, it has not been sufficiently strong relative to the rate of mutation [40], to generate a continuum of diversity throughout the genome. This means members of a 'metagenome strain' may differ substantially from each other particularly in rapidly evolving accessory regions and the subpopulation as a whole may descend from multiple cells but with a core genome that has descended from a single cell in the recent past. This is equivalent to the definition of 'lineage' in [29]. For ease, in the discussion below we will refer to strain in the metagenome context when properly we mean this looser definition of a strain as a genetically distinct subpopulation.

De novo assembly of genomes from short read metagenome sequences remains very challenging. Assemblies become fragmented for two reasons: firstly, low coverage genomes will fragment through chance occurrences where sequence coverage drops out, following Lander and Waterman statistics [17], secondly, if either intra or inter-genomic repeats are present then the assembly graphs used to represent possible sequence overlaps become very complex, and it is unclear which paths correspond to true genomes. Both of these issues are particularly problematic for metagenomes, where there can be a wide range of species abundances, and in a complex community a significant fraction of the species may be at low coverage. The first challenge can be addressed by sequencing more deeply. More difficult to address is the problem of repeats. Just as they do in isolate genome sequencing, intra-genomic repeats such as the 16S rRNA operon will lead to uncertainty in metagenomic assemblies, but if multiple closely related strains from the same species are present then they will possess potentially large regions of shared sequence. If the strain genomes are of comparable divergence to the reciprocal of the read length then very complex graphs will result, for typical short read sequencing (75-150bp) this would be strains at around 98-99.5% sequence identity. The result is that it is not possible to find long paths in the graph that can be unambiguously assembled into long contiguous sequence or contigs. For this reason metagenome assemblies for strain-diverse communities can comprise millions of contigs when made from short read data, with the added drawback that in the metagenomics context, we do not even know which contig derives from which species. For species that contain multiple very similar strains ($> 99.9\%$), then we expect better assemblies but the variants are then too far apart to be linked or phased by Illumina reads. In that case we may resolve the large-scale genome structure but not the sequences of the individual strains, which we will refer to as their haplotypes.

Metagenomic contig binning methods attempt to mitigate the problem introduced by standard metagenome sample processing approaches, wherein the origin of each sequence read is unknown. Contig binning works because contigs deriving from the same or similar genomes will share features that can be learnt without prior knowledge. These features can be sequence composition, but it is also possible to use per-sample coverage depths of contigs as a more powerful feature, if multiple samples are available from the same (or very similar) communities [2]. There are now numerous algorithms capable of using both coverage across samples and composition to automatically cluster contigs and determine from single-copy core gene (SCG) frequencies where the resulting bins are good quality metagenome assembled genomes (MAGs) [3, 13]. These tools enable genome bins to be extracted *de novo* from metagenomes, and are becoming crucial for studying unculturable organisms, contributing to many exciting discoveries, such as the description of the Candidate Phyla Radiation [9] or an improved understanding of the diversity of nitrogen fixers in the open ocean [14].

The resolution of genome binning though, is limited by the resolution of the assembler, with a typical maximum kmer length of around 100, the best case is that we can resolve to about 1% sequence divergence, so that bins correspond to something between a species and a strain. In the presence of strain diversity, those contigs that are shared across strains will become a consensus of the strains present, in the ideal situation their sequence would be that of the most abundant strain, but even this is not guaranteed. Contigs that are part of the accessory genome and present in a subset of strains may be successfully binned with the core genome, but they may not if they are too short or divergent in coverage. Consequently, if multiple strains are present in the assembly the MAGs that result from binning will be an imperfect composite of multiple strains.

Strains in a metagenome can exhibit variation in shared genes, such as insertions/deletions and single-nucleotide variants or SNVs, as well as in their accessory gene complements. Recently, we introduced DESMAN [32] to resolve subpopulations in MAGs using variant frequencies on contigs when multiple samples from a community are available. This is similar to contig binning using coverage but it can be viewed as a relaxed form of clustering closer to non-negative matrix factorisation, because each variant can appear in more than one subpopulation haplotype. Similar strategies had been proposed prior to DESMAN but using variant frequencies on reference genomes e.g. Lineages [29] and Constrains [27]. DESMAN and other earlier methods are all ‘linear mapping-based methods’ where metagenomic reads are mapped onto a linear sequence, either a reference or consensus contig. This has multiple drawbacks: firstly, the type of variant that can be represented is limited to changes at a single base; secondly, mapping onto a linear sequence can be challenging when there is variation present yielding unreliable results [19]; thirdly, it treats every variant as independent ignoring the co-occurrence of variants in reads, which is a powerful extra source of information when strain divergence is greater than the inverse of read length, when we would expect most reads to contain more than one variant. The last issue can be addressed by keeping track of which variants appear in which reads but that requires extra bookkeeping [18].

To address these limitations, we introduce a new method, STRONG (Strain Resolution ON Graphs), for analysing metagenome series when multiple samples are available either from the same microbial community e.g. longitudinal time-series or cross-sectional studies where the communities are similar enough to share a significant fraction of strains. STRONG can determine the number of ‘metagenome strains’ in a MAG formed from binning of a coassembly of all the samples, together with their sequences across multiple single-copy core genes, which we define as the strain haplotype, and the coverages of each strain in each sample. STRONG avoids the limitations of the variant-based approaches by resolving haplotypes directly on assembly graphs using a novel variational Bayesian algorithm, BayesPaths.

This graph-based approach allows more complex variant structure and incorporates read information. The usefulness of graphs for understanding microbial strains has been noted before, and efficient algorithms developed for querying complex graphs and extracting more complete representatives of MAGs in the presence of strain diversity [10]. STRONG, however, is the first time that graphs have been used in an automated workflow to actually decompose that strain diversity into haplotypes across multiple genes using multiple samples. We compare STRONG to the current state of the art, DESMAN, on synthetic microbial communities and a real metagenome time series from an anaerobic digester. In the former case we validate using the known genome sequences, and for the latter we compare abundant MAGs with haplotypes derived independently from Oxford Nanopore MinION long reads.

Results

STRONG pipeline

The detailed pipeline is described in the Methods but the key steps are summarised in Figure 1 and reiterated here. We start from multiple samples of the same community and jointly coassemble them with metaSPAdes, we save a high resolution graph (HRG) early in the assembly process that preserves all the variant information in the coassembly. The metaSPAdes assembly process then proceeds as normal and the resulting contigs are binned using CONCOCT. We annotate the single-copy core genes in the contigs, allowing us to identify a subset of bins as MAGs. A novel algorithm was then developed to map these SCG ORFs onto the HRG and extract the complete assembly subgraphs corresponding to the genes of interest (Methods - Relevant subgraph extraction). We obtained per sample unitig coverages on these subgraphs by threading reads directly onto them. These subgraphs were simplified with a noise filtering algorithm that used the MAG coverage depths, calculated as the length weighted average of the contigs assigned to that MAG. The simplified subgraphs contain all the information required for the BayesPaths algorithm (Methods - BayesPaths), that

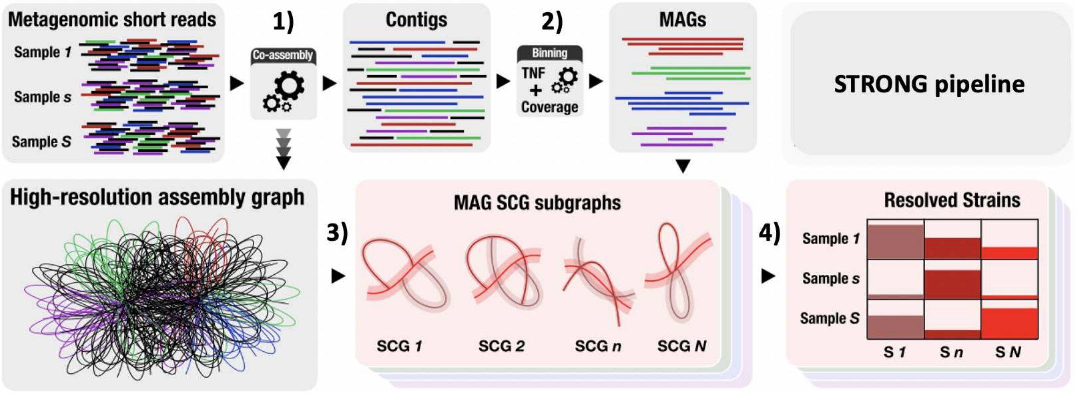

Figure 1. STRONG pipeline. This figure illustrates the principal steps in the STRONG pipeline (see Methods - STRONG Pipeline). Step 1) Co-assembly with metaSPAdes and storage of a high-resolution graph (HRG). Step 2) Contig binning with CONCOCT and annotation of single-copy core genes (SCGs). Step 3) Mapping of SCGs onto the HRG and extraction of individual SCG assembly graphs together with per-sample unitig coverages. Step 4) Joint solution of SCG assembly graphs from each MAG with BayesPaths to determine strain number, haplotypes and per-sample coverages.

simultaneously solves for the number of strains present, their coverage in each sample, and their sequences on the SCGs. SCGs from the same MAG are linked through the binning process and jointly solved in the strain resolution procedure to generate linked strain resolved sequences for each SCG. We will refer below to the SCG sequences for a given strain as its haplotype. The pipeline also applies DESMAN [32], to the same MAGs for comparative purposes, and will perform benchmarking if known genomes are available. It is important to note that some SCGs will be filtered during the BayesPaths procedure, see Methods, so that sequence inference is only performed on a subset in the final output.

Synthetic data sets

In order to provide an example metagenome data set with a known strain configuration for each species, we created a synthetic community comprised of 100 strains, with known genomes deriving from 45 species, with 20 species represented by a single strain, 10 with two strains, 5 with three, 5 with four and 5 species with five strains. We then generated four data sets from this community with the same total number of reads (150 million 2X150 bp) but increasing sample numbers (3, 5, 10 and 15 samples). This configuration, where most species have a single strain, might be an appropriate approximation to the human gut microbiome [38]. We denote these data sets Synth_S03, Synth_S05, Synth_S10 and Synth_S15. For each sample number, random species abundances were generated from a log-normal distribution, with strain proportions from a Dirichlet. Full details of the synthetic sequence generation are given in the Methods.

The STRONG pipeline was then applied to each of these data sets in turn. In Figure 2 we illustrate the STRONG output for a single gene, COG0532 ‘Translation initiation factor IF-2’ [37], from one MAG, Bin_55 of the ten sample synthetic data set, giving the resulting decomposition of the assembly subgraph into three strains. Noting that the strains were resolved in this MAG over 22 single-copy core genes simultaneously, and that for this 3.4 kbp

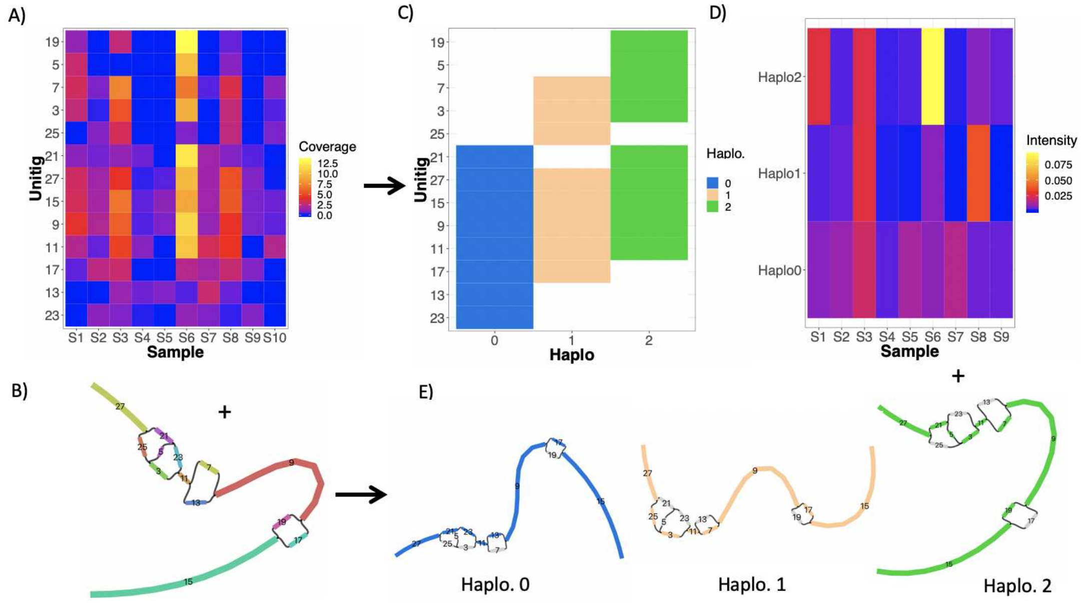

Figure 2. BayesPaths algorithm. This illustrates the BayesPaths algorithm for a single COG0532 from one MAG, Bin_55 of the ten sample synthetic data set. The algorithm predicted 3 strains. We show the input to the algorithm: A) the unitig coverages across samples plus B) the unitig graph without strain assignments. The outputs of the algorithm are shown in C) the assignments of haplotypes to each unitig, D) the strain intensities across samples, effectively coverage divided by read length (see Methods - BayesPaths), and E) unitig graphs for each haplotype with their most likely paths. This algorithm is explained in detail in the Methods - BayesPaths.

gene the haplotypes were found without errors.

For each of the four synthetic data sets we considered only MAGs which were assigned to species (see Methods) with at least two strains - 20, 21, 24 and 22 MAGs, from the Synth_S03, Synth_S05, Synth_S10 and Synth_S15 data sets respectively. For each MAG we mapped the predicted haplotypes for the optimal strain decomposition for both the STRONG pipeline and DESMAN algorithms onto the known reference strains. We then assigned each haplotype prediction to its best matching reference. The best such match was denoted 'Found'. If multiple predicted haplotypes matched to the same reference they were denoted as 'Repeated'. If a reference had no haplotype prediction that matched to it better than the other references, it was denoted as 'Not found'. For the aggregate across these MAGs we show the total number of such strains for each of the four data sets in Figure 3.

STRONG consistently outperforms DESMAN in terms of number of strains found, in total across all four samples it resolved 213 strains vs. 200 for DESMAN *i.e.* a 6.5% increase. It also had fewer 'Repeated' strains, 8 vs. 23: a reduction of 65%. The strains 'Found' were also reconstructed more accurately, the per base error rate for the BayesPaths reconstructions averaged across all MAGs and all data sets was just 0.052%, three times lower than that for DESMAN, 0.176%. This improvement was observed for all four data sets (see Table 1 and Figure 4). STRONG was more likely to predict the correct number of strains, doing so for 73% of MAGs summed across samples numbers versus 60% for DESMAN. It was also better at predicting the strain relative abundances. Regressing true abundance against predicted

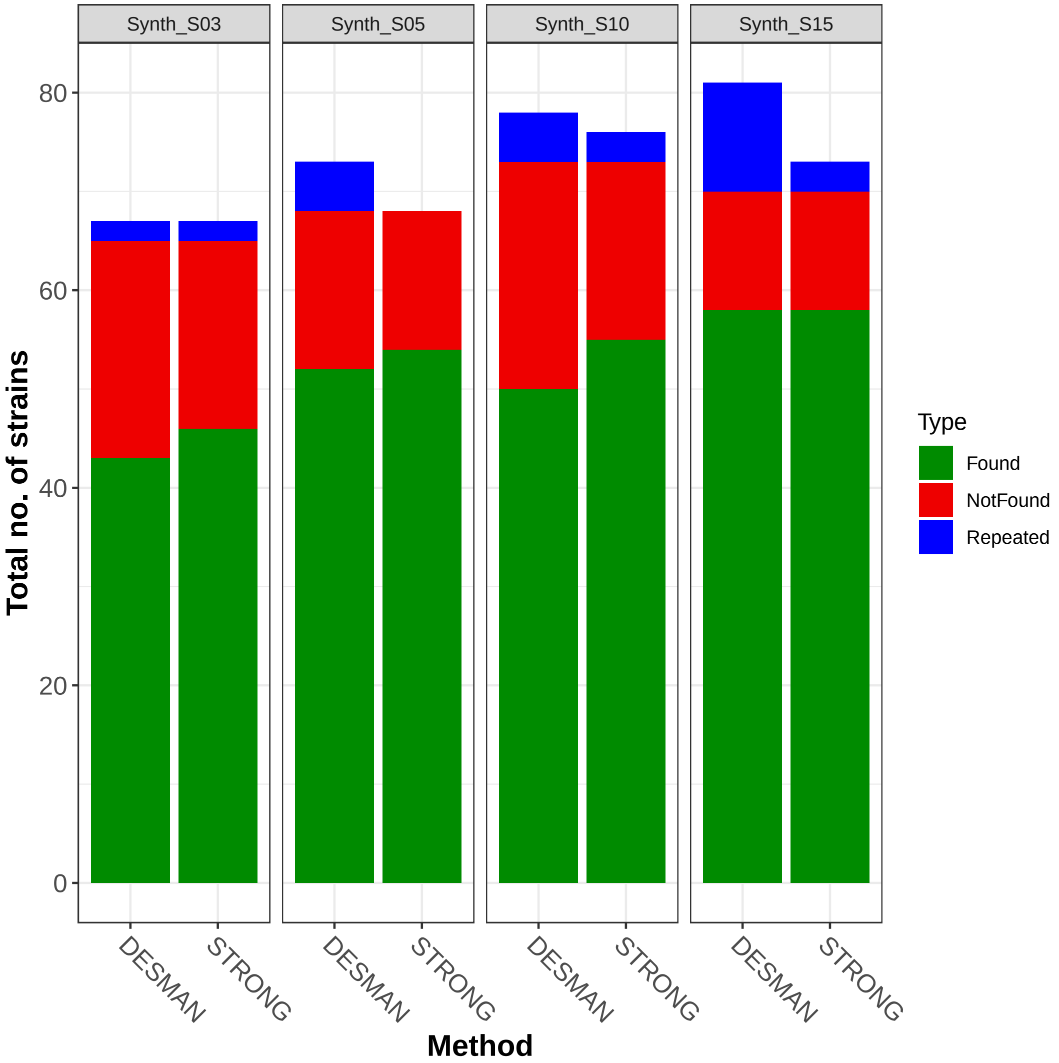

Figure 3. No. of strains resolved by STRONG and DESMAN algorithms in the synthetic community data sets. For MAGs with two or more strains we mapped haplotypes to the references and assigned each predicted haplotype to its best matching reference. The best such match was denoted 'Found'. If multiple haplotypes matched to the same reference they were denoted as 'Repeated'. If a reference had no predicted haplotypes matched to it, it was denoted as 'Not found'. The bars give the total numbers in each category summed over MAGs for the two methods (DESMAN and STRONG) and the panels results for the four different data sets with increasing number of samples (Synth_S03, Synth_S05, Synth_S10 and Synth_S15).

abundance gave an adjusted R^2 of 0.84 averaged across sample numbers for STRONG vs. 0.80 for DESMAN. When this was restricted to MAGs where the number of strains was correctly predicted, then both algorithms did better but STRONG still out performed DESMAN, with a mean R^2 of 0.98 compared to 0.93. Although the quantity varied across the four data sets, roughly 1/3 of the SCGs were filtered during the BayesPaths as outliers (see - Table 1).

The STRONG pipeline outperforms DESMAN, but it still misses strains that are present. In total across all MAGs and data sets, 63/276 *i.e.* 22.8%, of strains were missed by STRONG.

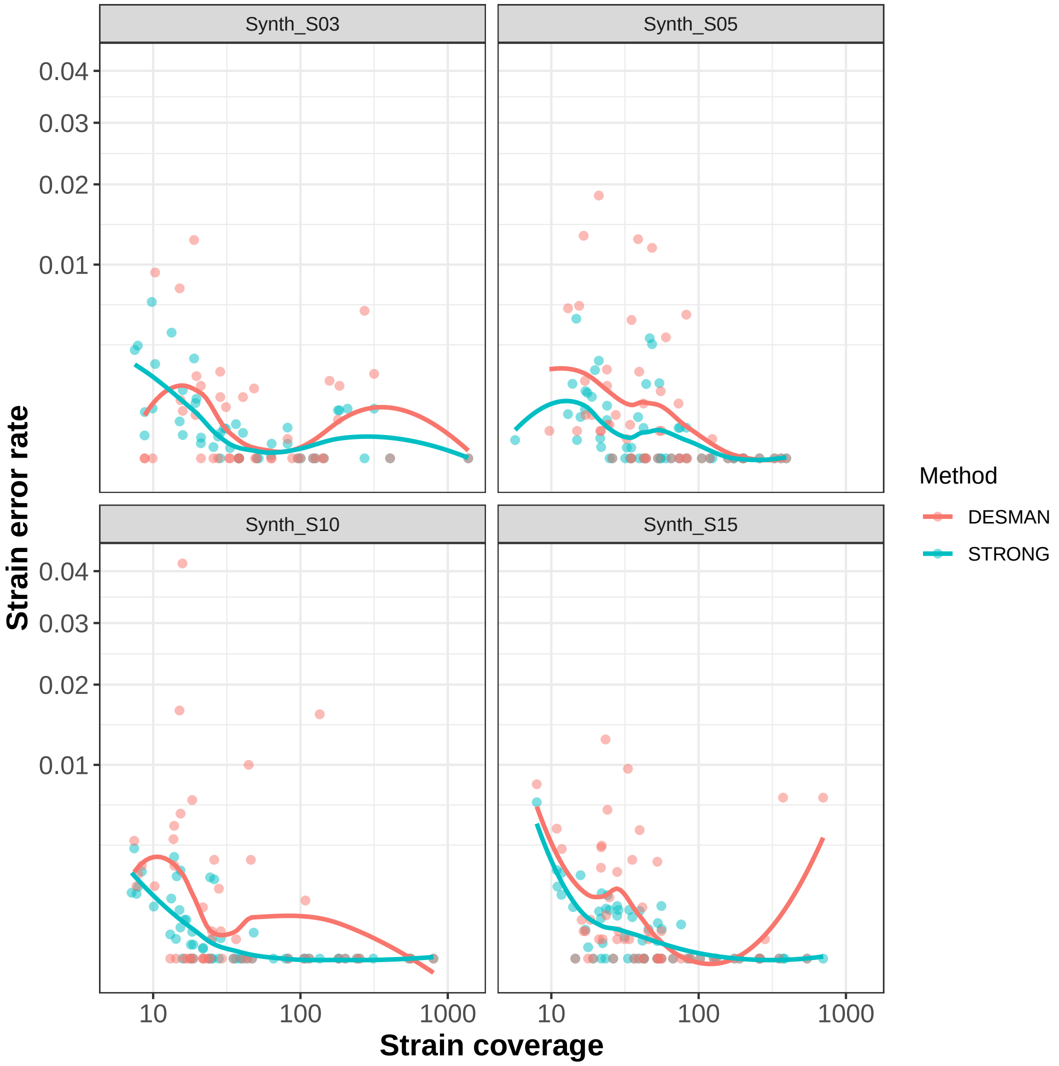

Figure 4. Error rates in strains found against coverage depth for STRONG and DESMAN algorithms in the synthetic community data sets. For the ‘Found’ strains we computed per base error rate to the matched reference, this is shown on the y-axis, against strain total coverage depth summed across samples on the x-axis, both axes are log transformed. The results are separated across methods (DESMAN and STRONG) and sample number in the synthetic community.

Some of these, 7 out of 63, were below the minimum coverage of detected strains (5.68), but most were not, suggesting that either they were not sufficiently divergent in terms of nucleotides or coverage profiles to be detected. Examination of phylogenetic trees for the haplotypes and reference genomes constructed using the SCGs revealed that in many cases ‘Not found’ strains had identical SCG haplotypes to those that were resolved.

The BayesPaths algorithm used to resolve strains in STRONG uses variational inference (see Methods - BayesPaths), an approximate Bayesian strategy [7]. This has the advantage of providing estimates of uncertainty in the inference of both the strain haplotypes and their abundances. The algorithm predicts the marginal probabilities that a given strain passes

Method	Data set	MAGs	#SCGs	#fSCGs	Found	Not F.	Rep.	Err	R^2	f^G
STRONG	Synth_S03	20	35.75	18.95	46	19	2	0.068	0.86 (0.99)	23/34 = 0.68
DESMAN					43	22	2	0.121	0.79 (0.93)	24/34 = 0.71
STRONG	Synth_S05	21	35.29	22.14	54	14	0	0.054	0.82 (0.98)	28/37 = 0.76
DESMAN					52	16	5	0.186	0.83 (0.99)	20/37 = 0.54
STRONG	Synth_S10	24	32.23	21.91	55	18	3	0.042	0.83 (0.99)	26/39 = 0.67
DESMAN					50	23	5	0.252	0.76 (0.95)	21/39 = 0.54
STRONG	Synth_S15	22	35.45	23.36	58	12	3	0.045	0.86 (0.98)	32/40 = 0.80
DESMAN					58	12	11	0.143	0.81 (0.87)	25/40 = 0.63

Table 1. Comparison of STRONG to DESMAN for strain reconstruction in the synthetic community data sets. Data set: Results are shown for the four different sample numbers. MAGs: The number of MAGs reconstructed with more than two reference strains. #SCGs: The average number of SCGs found in each MAG. #fSCGs The average number of SCGs after filtering in STRONG. Found: Number of reference strains that had a predicted strain that best matched it. Not F.: Number of reference strains that had no predicted strain with a closest match to it. Rep.: Number of reference strains with more than one best matching predicted strain. Err: The average error rate of the 'Found' strains in percentage base pairs. R^2: Correlation between predicted and actual strain relative proportions given as adjusted R^2, the figure in parentheses is when restricted to MAGs where the number of strains was correctly predicted. f^G: the fraction of MAGs where the number of strains was correctly inferred.

through a particular unitig. To provide a single sequence for the evaluation above and applications below we output the most likely path and hence sequence for each strain. However, we also calculate an estimate of path uncertainty by sampling many possible paths (default 100) consistent with the marginal distributions and calculate the average number of nodes that deviate from the most likely path, we refer to this as the divergence. For the 'Found' strains this correlates strongly with actual error rate to the reference strain (Pearson's correlation $r = 0.56$, $p < 2.2e - 16$ - see Figure S1). Thus the divergence is a useful prediction of uncertainty in the haplotype sequence inference, enabling us to estimate error rates in real data sets in the absence of known reference sequences. Roughly speaking, the expected per base error rate is 0.01 times the divergence, so that a strain divergence of 0.1 predicts a 0.1% error rate. In real data sets, the uncertainty estimates in the abundances are also useful, placing bounds on the abundance of individual strains in each sample.

In Table S3 we give approximate run times for each component of the STRONG pipeline on the synthetic community data sets, using 64 threads on a standard bioinformatics server (see Table S3). The BayesPaths step is the most time consuming part of the analysis (up to 36 hours), but it is still comparable to the initial coassembly. The only part of the pipeline with substantial memory requirements is the initial coassembly with metaSPAdes, the other steps are CPU limited.

Anaerobic digester time series

We next applied the STRONG pipeline to a real metagenomics time series, comprising ten samples taken at approximately 5 weekly intervals, from an industrial anaerobic digestion reactor (see Table S4 and Methods for details). This provides an evaluation community of intermediate complexity to test the pipeline's capability to resolve strains and reconstruct intraspecies dynamics. Each sample was sequenced on the NovaSeq platform with 2x150 bp reads at a mean depth of 11.63 Gbp. One sample was also run on a Nanopore MinION flow cell producing 43.78 Gbp of reads with a read N50 of 6,727 bp and a maximum length of 108 kbp.

CONCOCT binning produced 905 bins, of which 309 had 75% of SCGs present in single-copy, which we designate MAGs. In total 11 of these MAGs exhibited overlapping SCG graphs and were merged into 6 composite MAGs (see Methods - STRONG Pipeline), so that 304 MAGs were actually used in the strain decomposition. We calculated coverage depth per sample for each bin and then normalised by sample size to obtain a community profile at each time point. Overall the reactor exhibited a clear shift in community structure over time, despite consistent operating conditions, with sample time explaining 48% of the variation in community structure ($p = 0.001$ - Figure S2). Of the MAGs, 110 had an abundance that changed significantly over time (Bonferonni adjusted p-value < 0.05 from Pearson's correlation of log transformed normalised abundance) and these were evenly split between those that increased (55) or decreased in abundance (55).

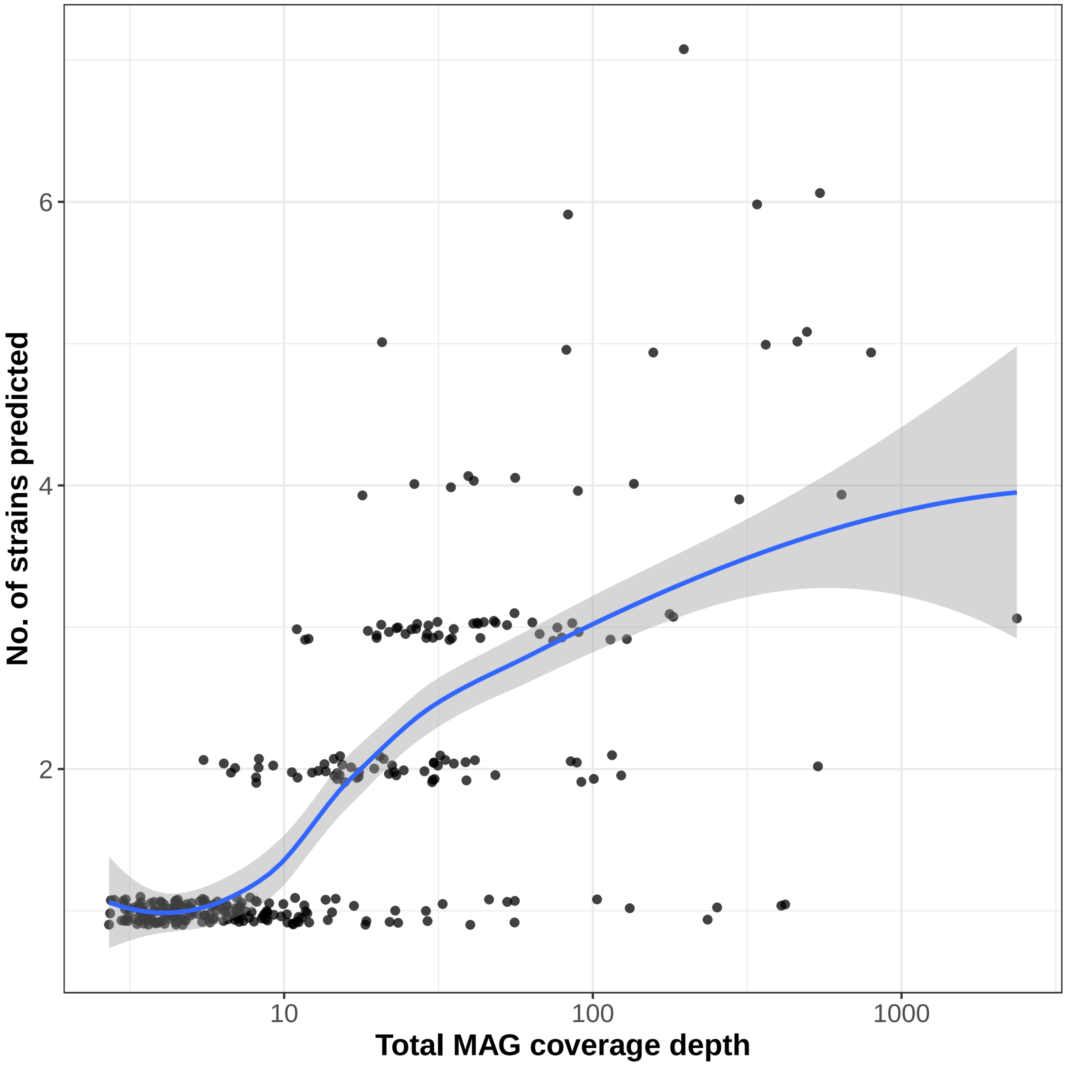

Figure 5. Number of strains resolved by STRONG against MAG coverage depth for the AD time series. Pearson's correlation between coverage depth and number of strains ($r = 0.36$, $p = 1.004e - 10$). The curve indicates a LOESS smoothing.

We used the STRONG algorithm to resolve strains in the 304 MAGs. This is a complex data set and running the complete pipeline took over 16 days, of which roughly 60% of the time was spent on the BayesPaths strain resolution (see Table S3). The number of strains found varied between 1 and 7, with a mean of 1.7, shown as a function of coverage depth in Figure 5. In total 121 (39.8%) of these MAGs had more than one strain, and there was a significant positive association between MAG coverage depth and number of strains ($r = 0.36$, $p = 1.004e - 10$), which is expected, as low coverage MAGs will be under-sampled. This correlation disappears though when we restrict to all MAGs with a coverage greater than thirty ($r = 0.19$, $p = 0.1023$). On average 20.9 SCGs were used after filtering for strain haplotype predictions.

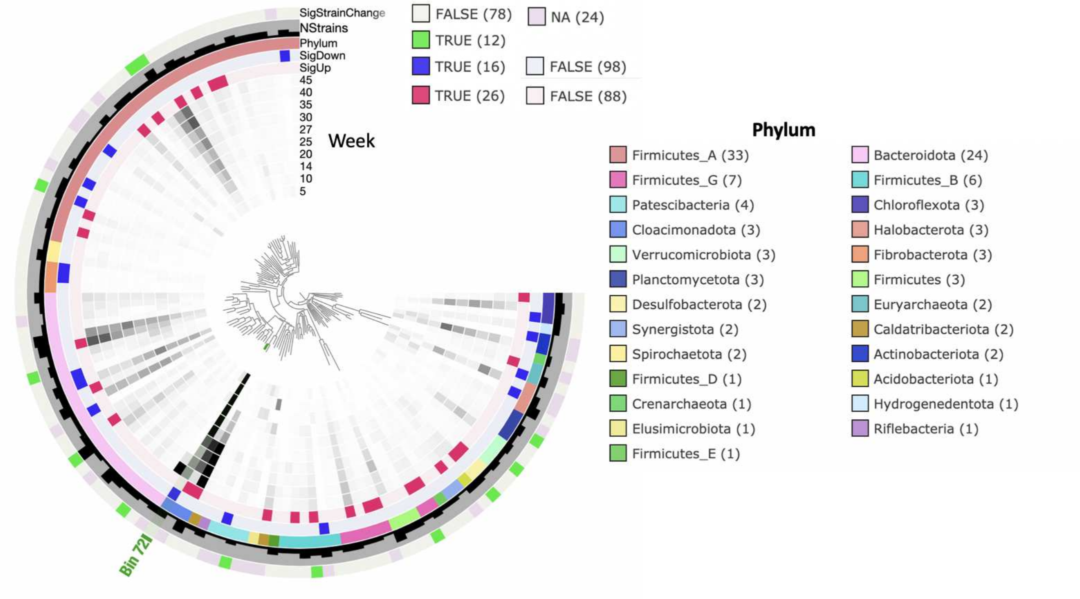

Figure 6. MAG summary for anaerobic digester time series. For the 114 MAGs with aggregate coverage > 20 we give their phylogeny constructed using concatenated marker genes together with their normalised coverages in the ten samples. We also indicate which MAGs significantly increased (SigUp) or decreased (SigDown) in total abundance (adjusted $p < 0.05$), their GTDB phylum assignment, no. of strains resolved by STRONG and whether the strain abundances changed significantly over time (adjusted $p < 0.05$) using permutation ANOVA (SigStrainChange).

For the 108 MAGs that had at least two strains with relative frequencies determined in five or more samples we used permutation ANOVA to determine whether strain proportions depended on sampling time. In total 13 of the MAGs had an adjusted p-value < 0.05 *i.e.* 12.0%. For these same MAGs 37 had a total coverage that changed significantly over time with an adjusted p-value < 0.05 *i.e.* 34.2%. Therefore the intra-species dynamics are more stable than inter-species, with strain proportions remaining fixed as the MAG coverages vary, this was true for 33 of the 37 MAGs that changed significantly in coverage.

In Figure 6, we use the Anvi'o program [16] to summarise information on phylogeny, taxonomy, normalised coverages in the ten samples, and whether the MAGs changed significantly in total abundance, together with the number of strains resolved by STRONG and if those strain relative proportions changed significantly with time. This was restricted to just those 114 MAGs with an aggregate coverage greater than twenty to simplify the diagram.

The Nanopore sequencing provides us with a means to directly test the validity of the STRONG haplotype reconstructions, at least for the most abundant MAGs. The most

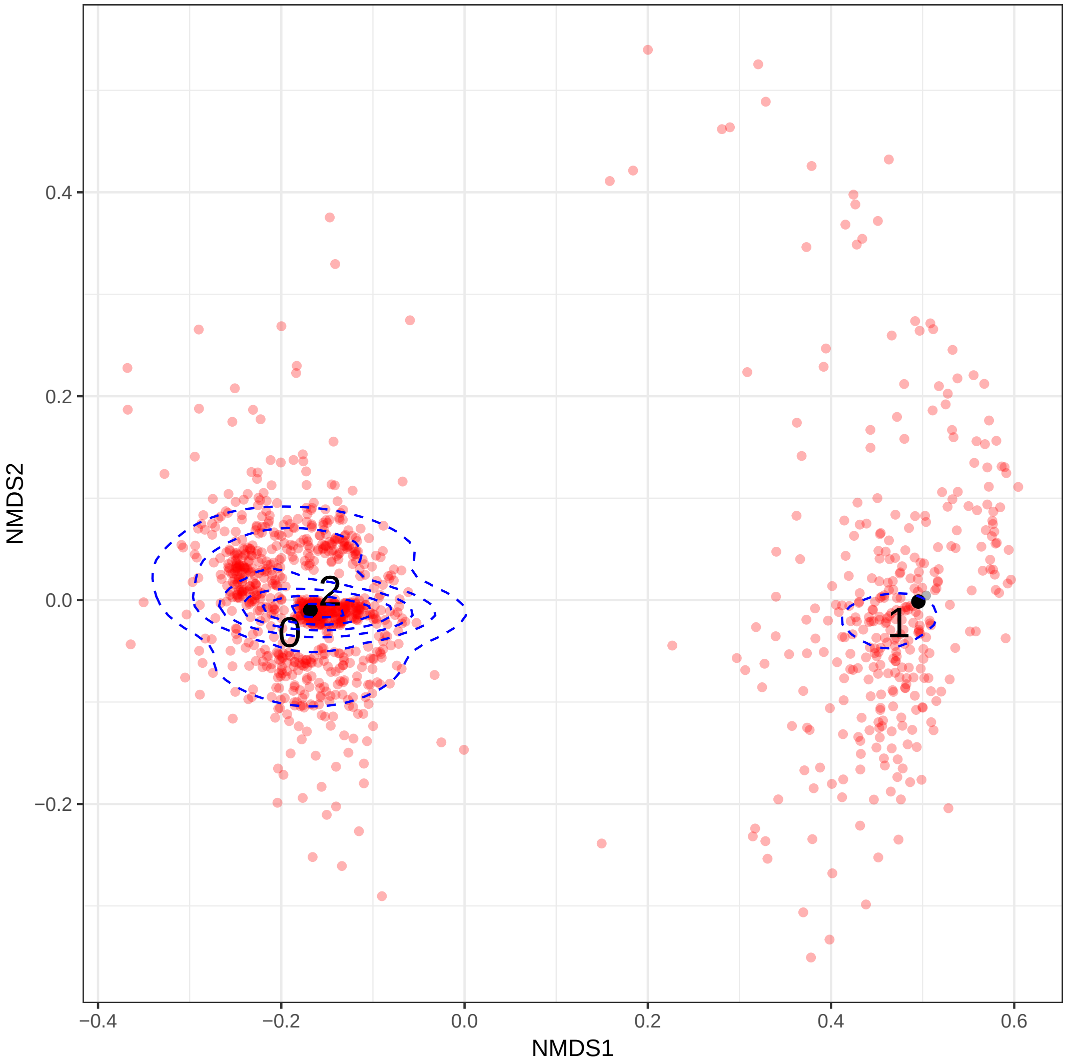

Figure 7. Comparison of Nanopore reads to STRONG prediction for COG0532 from Bin_72. Non-metric multidimensional scaling of Nanopore reads that mapped to COG0532 from Bin_72 of the anaerobic digester time series (red) together with the three haplotypes reconstructed from short reads by STRONG (black 0, 1 and 2). Haplotypes 0 and 2 were identical for COG0532. Distances were calculated as fractional Hamming distances (see text) on short read variant positions (see Methods - Nanopore Sequence Analysis). Blue dashed lines indicate read density contours.

abundant MAG, Bin_72, had an aggregate short read coverage depth of 2364.25, across all the samples. This MAG was assigned to the phylum Cloacimonadota using the GTDB taxonomy [11]. Interestingly, this is an example of a MAG which changes significantly in abundance, decreasing over time, (adjusted $p = 4.9e-05$) but where the proportions of the three strains predicted varied less dramatically ($R^2 = 0.35$ adjusted $p = 0.089$) - see Figure S5. We will focus on the longest SCG for which strains were resolved, COG0532, where the three strains are present in only two variants, haplotypes 0 and 2 being identical on this core gene. In Figure S4 we give the short read variant graph for this gene, which in this case is mostly simple bubbles, together with the assigned haplotypes. In fact, across the 18 SCGs used to decompose strains, haplotypes 0 and 1 were most similar with 99.7% nucleotide identity. These two strains had 99.4% and 99.1% identity with haplotype 2 respectively. That this pattern was not observed on COG0532 may suggest some recombination in the evolution of these organisms.

In Figure 7 we show for COG0532 both the Nanopore reads that map to this gene and the three haplotypes inferred by BayesPaths, as an Non-metric Multi-dimensional Scaling (NMDS) plot using fractional Hamming distances on the short read variant positions. These are defined as the Hamming distance between two reads but only on the intersecting variant positions and ignoring gaps. We then normalise by the number of such non-gap intersecting positions to give a distance between 0 and 1. The Nanopore reads are consistent with the inference of two variants on this gene, as there are two clear clusters observed, and the two modes of those clusters are close to those haplotypes. The variation around the modes is caused by the high error rate of the Nanopore reads.

In order to provide a quantitative comparison of the Nanopore reads and the STRONG predictions, we applied the EM algorithm defined in the Methods (Nanopore Sequence Analysis) on the 1,603 Nanopore reads mapping to this COG (cluster of orthologous groups). Examining the negative log-likelihood as a function of number of strains, it flattens at two strains (see Figure S3) and the two strains inferred exactly match (100% identity over 2,313 bps) haplotypes 0/2 and 1 respectively. Furthermore, STRONG in this sample predicted frequencies of 28.0% for haplotype 1. This closely matched the Nanopore haplotype frequencies for this strain of 27.6%. We also ran the Nanopore EM algorithm for all 18 filtered COGs in this bin separately. For the 11 COGs where more than one strain was predicted from the Nanopore reads, we compared the STRONG and Nanopore predictions. For haplotypes 0, 1 and 2 exact matches were found for 6, 7 and 4 SCGs respectively with average nucleotide identities across all genes of 99.89%, 99.89% and 99.82%.

For lower coverage MAGs we generally obtain a reasonable correspondence between the STRONG haplotypes and Nanopore predictions. In most cases the number of strains is comparable between the two, although the accuracy of matches reduces with decreased Nanopore read counts, as we might expect. As an example, in Figure S7 we compare Nanopore reads with the five STRONG haplotypes from COG0072 of Bin_846, a Firmicutes MAG in the AD time series. The most abundant Nanopore mode clearly matches STRONG haplotype 4, the most abundant strain in this sample, and there is also some support for haplotypes 0 and 2. There is less evidence for strains 1 and 3, but these are low abundance in this sample (see Figure S9). This is confirmed from the EM algorithm applied to the Nanopore reads matching this gene, where we would predict four Nanopore haplotypes (Figure S6). Comparing these 4 Nanopore strains to the STRONG predictions we find that three closely match: Nanopore haplotype 0 matched best to STRONG strain 4 with 98.8% nucleotide identity, Nanopore haplotype 1 to STRONG 4 with 99.9% identity and Nanopore haplotype 2 to STRONG strain 0 with 99.7% identity. There is also a correspondence in relative abundance, with the most abundant Nanopore haplotype 1 recruiting 82% of the reads vs 74% relative frequency for the corresponding strain haplotype from STRONG.

Discussion

We have demonstrated that on synthetic data the STRONG pipeline and the BayesPaths algorithm are able to accurately infer strain sequences on the SCGs and abundances across samples. Performance does improve with increasing sample number in terms of the number of strains resolved, but reassuringly even when only a small number of samples are available we are still able to accurately predict (with 0.068% per base error rate) strains, and when ten or more samples are available we obtain error rates below 0.05% *i.e.* 1 error in every 2000 bps from short read data. This is better performance than the state of the art, and sufficiently accurate for high resolution phylogenetics. Strains are resolved more accurately as they increase in coverage (see Figure 4), and in fact, when coverages exceed twenty fold we can resolve strains very reliably, with just 0.011% error rate averaged across strains in the ten sample synthetic data set. We believe therefore that this pipeline will be useful whenever high quality *de novo* strains are required from metagenome short read time series.

This is to our knowledge the first algorithm capable of constructing strains from metagenomes using assembly graphs from multi-sample coassemblies. Graph-based haplotype resolution has been applied to viruses [4] and for eukaryotic transcripts [5,6], but ours is the first algorithm to resolve strains across multiple gene subgraphs connected through a contig binning procedure. The BayesPaths algorithm is also a substantial algorithmic advance enabling coverage across multiple samples to be incorporated into a rigorous Bayesian procedure that gives uncertainties in both the paths (*i.e.* the sequences) and the strain abundances. This algorithm could be utilised outside of the actual STRONG pipeline in other application areas, for example for finding viral haplotypes.

In addition, to the new strain resolution algorithm, BayesPaths, STRONG incorporates a number of useful tools for large-scale variant graph processing, in particular, the tools for extraction of subgraphs that correspond to individual coding genes and the spades-gsimplifier tool for error correction on those graphs. These can be applied to any graph in the GFA format, and could therefore find applicability outside of the context of our pipeline. This also means that in the future we could add alternative choices for the coaseembly step, for instance replacing metaSPAdes with MEGAHIT [25]. Similarly, we plan to add support for alternative binners to CONCOCT.

Currently, we are restricted to core genes that are single-copy and shared across all strains in a MAG. We can in theory use any such genes, so if a particular MAG is of interest the pipeline could be run with a larger set of COGs that are SCGs for that MAG. There would be a cost in terms of increased running time, which will increase with more genes and unitigs in a roughly linear fashion.

The analysis of a time series from an anaerobic digestor illustrates the practicality of our pipeline on a realistically sized data set. We should note though that to resolve strains on these 304 MAGs took nearly 10 days using 64 threads on a standard bioinformatics server (see Table S3. The AD analysis also demonstrates the importance of strain dynamics in a real microbial community with nearly 40% of MAGs exhibiting strain variation, but this variation was relatively stable compared to the MAG dynamics themselves. If strains are functionally redundant to one another we would expect significant neutral fluctuations over time. Therefore this could be evidence for intra-species niche partitioning.

In general, we found a good correspondence between haplotypes inferred from Nanopore reads and the STRONG predictions in the AD data set. For the most abundant MAG, Bin_72, they matched very closely. In addition, the relative abundances of strains were consistent across the two sequencing technologies, despite the use of different DNA extraction protocols, and the different biases inherent in library preparation and sequencing platforms. These technical elements in the data generation process are known to introduce bias at the species level [12], but our findings suggest that intraspecies abundance may generally be robust against such biases, which makes sense in that all the strains of a species will have similar physical properties and genomic traits.

STRONG is an effective strategy to *de novo* resolve subpopulations at high phylogenetic resolution within MAGs, but as discussed in the Introduction, it is important to add the caveat that the haplotype sequences obtained are not equivalent to those from sequencing cultured isolates, where we can identify the resulting genome with a single organism present in the original community. The metagenome strains, in the best case, will correspond to different modal sequences of the taget species, about which substantial unresolved variation may exist. They will correspond to peaks in the probability distribution of all possible sequence configurations, and as such will provide important insights into the naturally occurring variation, but there remains the question of how to identify and quantify the unresolved variation surrounding those peaks. In the worst case, when STRONG is applied to rapidly recombining microbes, such as those found in the oceans [30], the resulting sequences may not even be real in the sense of characterising any true individual. An additional unaddressed question is how to determine when this has occurred, for now we would simply urge caution when using STRONG in cross-sectional studies of rapidly evolving microbes, and suggest that the term ‘metagenome strain’ or ‘metagenome haplotype’ be used when referring to the output sequences. The same caveat does of course apply to any current purely bioinformatics strategy for *de novo* resolution of genomes from metagenomes. Even if a single sample is used for binning and there are no subpopulations, the resulting MAG is still a composite and not a strain in the traditional microbiological sense [39].

An obvious extension of our algorithm would be to resolve the accessory genome into strain genomes. This could be done on a per gene basis by relaxing the requirement that every strain passes through every gene, but an approach that incorporates the path structure in the full metagenomic assembly would be more powerful. Use of the full assembly may be possible in an efficient manner by factorising the variational approximation on a per gene basis and allowing the solutions for one gene to depend on the expectations across their neighbours. Or it may be that more computationally tractable versions of the algorithm can be developed that will scale to larger graphs. In any case the issues discussed above of our inferred ‘strains’ containing unresolved variation will become more pertinent when we extend our algorithm to the full genome, and it will be necessary to consider not just the most likely genome associated with a subpopulation but also its variants.

In the future we also plan to directly incorporate long read information into the strain resolution rather than just using it for validation. It was encouraging therefore to see the correspondence in strain frequencies between the two approaches. We are confident that in the near future, through the combination of long reads with methods similar to those we have introduced in STRONG, that complete metagenome *de novo* strain resolution will become a realistic possibility.

Conclusion

We have introduced a complete bioinformatics pipeline, STrain Resolution ON assembly Graphs (STRONG), that is capable of extracting single-copy core gene variant graphs from short read metagenome coassemblies for individual metagenome assembled genomes (MAGs). We demonstrated how these graphs and associated per-sample unitig coverages can be used in a novel Bayesian algorithm, BayesPaths, to find MAG strain number, haplotypes and abundances. This approach achieves superior accuracy to variant based methods on synthetic communities and predictions on real data that match those from long Nanopore long reads.

STRONG is freely available from `https://github.com/chrisquince/STRONG`.

Methods

Synthetic data set generation

The *in silico* synthetic communities were generated by first downloading a list of complete bacterial genomes from the NCBI and selecting species with multiple strains present. Genomes were restricted to those that were full genome projects, possessed at least 35 of 36 single-copy core genes (SCGs) identified in [3], and with relatively few contigs (< 5) in the assemblies. Communities were created by specifying species from this list and the number of strains desired. The strains selected were then chosen at random from the candidates for each species, with the extra restrictions that all strains in a species were at least 0.05% and no more than 5% nucleotide divergent on the SCGs from any other strain in the species. This corresponds to a minimum divergence of approximately 15 nucleotides over the roughly 30 kbp region formed by summing the SCGs. The genomes used are given in Tables S1 and S2.

Each species indexed i was then given an abundance, $y_{i,s}$, in each sample, $s = 1, \ldots, S$, which was drawn from a lognormal distribution with a species dependent mean and standard deviation, themselves drawn from a normal and gamma distribution respectively:

$$\log(y_{i,s}) \sim N(\mu_i, \sigma_i)$$

where:

$$\mu_i \sim N(\mu_p, \sigma_p)$$

and:

$$\sigma_i \sim \mathrm{Gamma}(k_p, \theta_p).$$

For all four community configurations — S equal to 3, 5, 10 and 15 — we used $\mu_p = 1$, $\sigma_p = 0.125$, $k_p = 1$ and $\theta_p = 1$. The species abundances were then normalised to one ($y'_{i,s} = y_{i,s}/\sum_i y_{i,s}$). For each strain within a species its proportion in each sample was then drawn from a Dirichlet:

$$\rho_{g,s} \sim \mathrm{Dirichlet}(\alpha) \tag{1}$$

with $\alpha = 1$.

This allowed us to specify a copy number for each genome g in species i in each sample as $y'_{i,s}\rho_{g,s}$. We then generated 150 million paired-end 2x150 bp reads in total across all samples with Illumina HiSeq error distributions using the ART read simulator [21]. The code for the synthetic community generation is available from `https://github.com/chrisquince/STRONG_Sim`.

Synthetic data set evaluation

We can determine which contig derived from which reference genome by considering the simulated reads that map onto it. We know which reference each of these came from, enabling us to assign a contig to a genome as that which a majority of its reads derive from. We can then assign each MAG generated by STRONG to a reference species as the one which the majority of its contig's derive from weighted by the contig length.

Anaerobic digester sampling and sequencing

AD sample collection

We obtained ten samples from a farm anaerobic digestion bioreactor across a period of approximately one year. The sampling times, metadata and accession numbers are given in Table S4. The reactor was fed on a mixture of slurry, whey and crop residues, and operated between 35-40°C, with mechanical stirring. Biomass samples were taken directly from the AD reactor by the facility operators and shipped in ice-cooled containers to the University of Warwick. Upon receipt, they were stored at 4°C and then sampled into several 1-5mL aliquots within a few days. DNA was usually extracted from these aliquots immediately but some were first stored in a -80°C freezer until subsequent thawing and extraction.

AD short read sequencing

DNA extraction was performed using the Qiagen Powersoil extraction kit following the manufacturer's protocol. DNA samples were sequenced by Novogene using the NovaSeq platform with 2x150 bp reads at a mean depth of 11.63 Gbp.

AD long read sequencing

Anaerobic digester samples were stored in 1.8 mL Cryovials at -80°C. Samples were defrosted at 4°C overnight prior to DNA extraction. DNA was extracted from a starting mass of 250 mg of anaerobic digester sludge using the MP Biomedical$^{\text{TM}}$ FastDNA$^{\text{TM}}$ SPIN Kit for Soil (cat no: 116560200) and a modified manufacturers protocol. Defrosted samples were homogenised by pipetting and then transferred to a MP bio$^{\text{TM}}$ lysing matrix E tube (cat no: 116914050-CF). Samples were resuspended in 938 μL of Sodium phosphate buffer (cat no: 116560205).

Preliminary cell lysis was undertaken using lysozyme at a final concentration of 200 $ng/\mu L$ and 20 μL of Molzyme Bug Lysis^{TM} solution. Samples were mixed by inversion and incubated at 37°C for 30 min on a shaking incubator (< 100 rpm). Lysozyme was inactivated by adding 122 μL of MP bio MT buffer and mixing by inversion. Samples were then mechanically lysed in a VelociRuptor V2 bead beating machine (cat no: SLS1401) at 5 m/s for 20 seconds then placed on ice for five minutes.

Samples were centrifuged at 14000 g for five minutes to pellet the particulate matter and the supernatant was transferred to a new 1.5 mL microfuge tube. Proteins were precipitated from the crude lysate by adding 250 μL of PPS$^{\text{TM}}$ (cat no: 116560203) and then mixing by inversion. Precipitated proteins were pelleted for five minutes at 14000 g and the supernatant was transferred to 1000 μL of pre mixed DNA binding matrix solution (cat no: 116540408). Samples were mixed by inversion for two minutes.

DNA binding matrix beads were recovered using the MP bio$^{\text{TM}}$ spin filter (cat no: 116560210) and manufacturer based spin protocol. The binding matrix was washed of impurities by complete resuspension in 500 μL of SEWS-M solution (cat no: 116540405) and centrifuged at 14000 g for five minutes. The DNA binding matrix was then washed for a second time by resuspension in 500 μL of 80% EtOH followed by centrifugation at 14000 g for five minutes. Flow though was discarded and centrifuged at 14000 g for two minutes to remove residual EtOH. The binding matrix was left to air dry for 2 minutes then DNA was eluted using 100 μL of DES elution buffer at 56°C. Elute was collected by centrifugation at 14000 g for 5 minutes and stored at 4°C prior to library preparation. Eluted DNA concentration was estimated using a Qubit 4^{TM} fluorometer with the dsDNA Broad Range sensitivity assay kit (cat no: Q32853). 260:280 and 260:230 purity ratios were quantified using a Nanodrop^{TM} 2000.

A 1x SPRI clean up procedure was undertaken prior to library construction to further reduce contaminant carry over. Input DNA was standardised to 1.2 μg in 48 μL of H_2O using a qubit 4^{TM} fluorometer and dsDNA 1x High Sensitivity assay kit (cat no: Q33231). Library preparation was undertaken using the Oxford Nanopore© Ligation Sequencing Kit (SQK-LSK109) with minor modifications to the manufacturer protocol. The FFPE/End repair incubation step was extended to 30 min at 20°C and 30 min at 65°C, while DNA was eluted from SPRI beads at 37°C for 30 min with gentle agitation. The SQK-LSK109 long fragment buffer was used to ensure removal of non-ligated adaptor units and reduce short fragment carryover into the final sequencing library. The final library DNA concentration was standardised to 250 ng in 12 μL of EB using a qubit 4^{TM} fluorometer and dsDNA 1 x High Sensitivity assay kit.

Sequencing was undertaken for 72 hours on an Oxford Nanopore© R 9.4.1 (FLO-MIN106) flow cell with 1489 active pores. DNA was left to tether for 1 hour prior to commencing sequencing. The flow cell and sequencing reaction was controlled by a MinION^{TM} MKII device and the GUI MinKNOW V. 19.12.5. ATP refuelling was undertaken every 18 hours with 75 μL of flush buffer (FB). Post Hoc basecalling was undertaken using Guppy V. 3.5.1 and the high accuracy configuration (HAC) mode.

STRONG pipeline

STRONG processes co-assembly graph regions for multiple metagenomic datasets in order to simultaneously infer the composition of closely related strains for a particular MAG and their core gene sequences. Here, we provide an overview of STRONG. We start from a series of S related metagenomic samples, e.g. samples of the same (or highly similar) microbial community taken at different time points or from different locations.

The Snakemake based pipeline begins with the recovery of metagenome-assembled genomes (MAGs) [22]. We perform co-assembly of all available data with the metaSPAdes assembler [28], and then bin the contigs obtained by composition and coverage profiles across all available samples with CONCOCT [3]. Each bin is then analyzed for completeness and contamination based on single-copy core genes, and poor quality bins are discarded. The default criterion is that a MAG requires greater than or equal to 75% of the SCGs in a single copy. While we currently focus on this combination of software tools, in principle we could use any other software or pipeline for MAG recovery, e.g. we could use MEGAHIT as the primary assembler [25] or alternative binning tools or their combination. For each MAG we then extract the full or partial sequences of the core genes that we further refer to as single-copy core gene (SCG) sequences.

The final coassembly graph produced by metaSPAdes cannot be used for strain resolution because, as with other modern assembly pipelines, variants between closely related strains will be removed during the graph simplification process. Instead, we output the initial graph for the final K-mer length used in the (potentially) iterative assembly following processing by a custom executable — spades-gsimplifier based on the SPAdes codebase — to remove likely erroneous edges using a ‘mild’ coverage threshold and a tip clipping procedure. We refer to the resulting graph as a high-resolution assembly graph or HRAG.

The graph edges are then annotated with their corresponding sequence coverage profiles across all available samples. As is typical in de Bruijn graph analysis, the coverage values are given in terms of the k-mer rather than nucleotide coverage. Profile computation is performed by a second tool for aligning reads onto the HRAG: unitig-coverage. The potential advantage of this approach in comparison to estimation based on k-mer multiplicity, is that it can correctly handle the results of any bubble removal procedure that we might want to add to the preliminary simplification phase in future.

For each detected SCG sequence (across all MAGs) we next try to identify the subgraph of the HRAG encoding the complete sequences of all variants of the gene across all strains represented by the MAG. The procedure is described in more detail in the next section. During testing we faced two types of problems here: (1) related strains might end up in different MAGs and (2) some subgraphs might consist of fragments corresponding to several different species. We take several steps to mitigate those problems. Firstly, we compare SCG graphs between all bins, not just MAGs. If an SCG graph shares unitigs between bins, then it is flagged as overlapping. If multiple SCG graphs between MAGs (> 10) overlap then we merge those MAGs, combining all graphs and processing them for strains together. Following merging any MAG SCG graphs with overlaps remaining are filtered out and not used in the strain resolution.

After MAG merging and COG subgraph filtering we process the remaining MAGs one by one. Before the core 'decomposition' procedure is launched on the set of SCG subgraphs corresponding to a particular MAG, they are subjected to a second round of simplification, parameterised by the mean coverage of the MAG, to filter nodes that are likely to be noise again by the spades-gsimplifier program. This module is described in more detail below. The resulting set of simplified SCGs of the HRAG for a MAG are then passed to the core graph decomposition procedure, which uses the graph structure constraints, along with coverage profiles associated with unitig nodes, to simultaneously predict: the number of strains making up the population represented by the MAG; their coverage depths across the samples; paths corresponding to each strain within every subgraph (each path encodes a sequence of the particular SCG instance).

A fraction of the SCGs in a MAG may properly derive from other organisms due to the possibility of incorrect binning *i.e contamination*. In fact, the default 75% single-copy criterion allows up to 25% contamination. In addition, the subgraph extraction is not always perfect. We therefore add an extra level of filtering to the BayesPaths algorithm, iteratively running the program for all SCGs, but then filtering those with mean error rates, defined by Equation 18, that exceed a default of 2.5 times the median deviation from the median gene error. Filtering on the median deviation is in general a robust strategy for identifying outliers. As a result of this filtering the pipeline only infers strain sequences on a subset of the input SCGs. We have found, however, that the number of SCGs for which strain haplotypes are inferred is sufficient for phylogenetics.

Relevant subgraph extraction

Provided with the predicted (partial) gene sequence, $\bar{T}$, and the upper bound on the length of the coding sequence, L, defined as $3\alpha\langle\bar{U}_n\rangle$ where $\langle\bar{U}_n\rangle$ is the average length in amino acids of that SCG in the COG database, and $\alpha = 1.5$ by default. The procedure for relevant HRAG subgraph extraction involves the following steps. First, the sequence $\bar{T}$ is split into two halves, $\bar{T}'$ and $\bar{T}''$, keeping the correct frame (both $\bar{T}'$ and $\bar{T}''$ are forced to divide by 3). $\bar{T}'$ and $\bar{T}''$ are then processed independently. Without loss of generality we describe the processing of $\bar{T}'$:

1. Identify the path $\mathcal{P}$ corresponding to $\bar{T}'$ in the HRAG. We denote its length as L_P.

2. Launch a graph search of the stop codons to the right (left) of the rightmost (leftmost) position of $\bar{T}'$ ($\bar{T}''$). The stop codon search is frame aware and is performed by a depth-first search (DFS) on the graph in which each vertex corresponds to a pair of the HRAG position and the partial sequence of the last traversed codon [1]. Vertices of this 'state graph' are naturally connected following the HRAG constraints. The search is cut off whenever a vertex with a frame state encoding a stop codon sequence is identified. Several stop codons can be identified within the same HRAG edge sequence in 'different frames', moreover the procedure correctly identifies all stop codons even if the graph contains cycles (although such subgraphs may be ignored in later stages of the pipeline).

3. The 'backward' search of the stop codons 'to the left' is actually implemented as a 'forward' search of the complementary sequences from the complementary position in the graph. Note that, as in classic ORF analysis, while the identified positions of the stop codons 'to the right' correspond to putative ends of the coding sequences for some of the variants of the analyzed gene, positions of the stop codons 'to the left' only provide the likely boundary for where the coding sequence can start. In particular, left stop codons are likely to fall within the coding sequence of the neighbouring gene (in a different frame). Actual start codons are thus likely to lie somewhere on the path (with sequence length divisible by 3) between one of the 'left' stop codons and one of the 'right' stop codons. For reasons of simplicity, further analysis of edges on the paths between left (right) stop codons ignores the constraint of divisibility by 3.

4. After the sets of 'left' and 'right' stop codon positions are identified along with the shortest distances between them and the $\bar{T}'$ path, we attempt to gather the relevant subgraph given by the union of edges lying on some path of a constrained length (see further) between some pair of left and right stop codons. First, for each pair (s, t) of the left and right stop codon positions we compute the maximal length of the paths that we want to consider $L_{s,t}$ as $L_P +$ min dist from s to start of $\mathcal{P} +$ min dist from end of $\mathcal{P}$ to t. The edge $e = (v, w)$ is considered relevant if there exists a pair of left (right) stop codon positions (s', t') such that the edge e lies on the path of length not exceeding $L_{s,t}$ between s' and t', which is equivalent to checking that $\text{min\_dist}(s', v) + \text{length}(e) + \text{min\_dist}(w, t') < L_{s,t}$. To allow for efficient checks of the shortest distances we precompute them by launching the Dijkstra algorithm from all left (right) stop codon positions in the forward (backward) direction [2].

5. We then exclude from the set of relevant edges the edges that are too far from any putative (right) stop codon to be a part of any COG instance. In particular, we exclude any edge $e = (v, w)$, such that the minimal distance from vertex w to any of the right stop codon positions exceeds L.

[1] Due to the properties of the procedure and the fact that it deals with DBGs, the actual implementation encodes frame state as an integer [0,2] rather than the string of last partially traversed codon.

[2] Actually Dijkstra runs are initiated from the ends/starts of corresponding edges and the distances are later corrected.

6. After the sets of the graph edges potentially encoding the gene sequence are gathered for $\bar{T}'$ and $\bar{T}''$ the union of the two sets, $\mathcal{ER}$, is then taken and augmented by the edges, connecting the 'inner' $\mathcal{ER}$ vertices (vertices which have at least one outgoing and at least one incoming edge in the $\mathcal{ER}$) to the rest of the graph.

Initial splitting of $\bar{T}$ into $\bar{T}'$ and $\bar{T}''$ is required to detect relevant stop codons which are not reachable from the last position of $\bar{T}$ in HRAG (or from which the first position of $\bar{T}$ in HRAG can not be reached). In addition to the resulting component in gfa format, we also store the positions of the putative stop codons, and ids of edges connecting the component to the rest of the graph (further referred to as 'sinks' and 'sources').

Subgraph simplification

While processing SCG subgraphs from a particular MAG we use the available information on the coverage of the MAG in the dataset. In particular, we set up the simplification module to remove tips (a node with no successors) below a certain length and edges with coverage below a fraction of the total coverage across all samples. If a tip is not removed it is labelled as a 'dead-end' to distinguish it from potential connections to the rest of the graph.

While simplifying a COG subgraph, edges connecting it to the rest of the assembly graph should be handled with care (in particular, they should be excluded from the set of potential tips). This is because in the BayesPaths algorithm they form potential sources and sinks of the possible haplotype paths. Moreover, during the simplification the graph changes, and such edges might become part of longer edges. Since we are interested in which dead-ends of the component do, and do not lead to the rest of the graph, the output contains the up-to-date set of connections of the simplified component to the rest of the graph.

We now briefly describe the implemented procedures based on 'relative coverage' criteria. Amongst other procedures for erroneous edge removal SPAdes implements a procedure considering the ratio of the edge coverage to the adjoining coverage of edges adjacent to it. We define an edge e as 'predominated' by vertex v incident to it if there is edge e_1 outgoing from v and edge e_2 incoming to v whose coverages exceed the coverage of e at least by a factor of α (by default equal to 5). Short edges (shorter than $k + \epsilon$) predominated by both vertices incident to them are then removed from the graph. Erroneous graph elements in high genomic coverage graph regions often form subgraphs of three or more erroneous edges. SPAdes implements a procedure for search (and subsequent removal) of subgraphs limited by a set of predominated edges. Starting from a particular edge (v, w) predominated by vertex v, the graph is traversed from vertex w breadth-first without taking into account the edge directions. If the vertex considered at the moment predominates the edge by which it was entered, the edges incident to it are not added to the traversal. The standard limitation of erroneous edge lengths naturally transforms into a condition of maximum length of the path between the vertices of the traversed subgraph. A limit on the maximum total length of its edges is additionally introduced.

BayesPaths

The model

We define an assembly graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ as a collection of unitig sequence vertices $\mathcal{V} = 1, \ldots, V$ and directed edges $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$. Each edge defines an overlap and comprises a pair of vertices and directions $(u^d \rightarrow v^d) \in \mathcal{E}$ where $d \in \{+, -\}$ and indicates whether the overlap occurs between the sequence $(+)$ or its reverse complement $(-)$. We define:

- Counts $x_{v,s}$ for each unitig $v = 1, \ldots, V$ in sample $s = 1, \ldots, S$
- Paths for strain $g = 1, \ldots, G$ defined by $\eta^g_{u,v} \in 0, 1$ indicating whether strain g passes through that edge in the graph
- Flow of strain g through unitig v, $\phi^{g+}_v = \sum_{u \in A(v)} \eta^g_{u,v}$ and $\phi^{g-}_v = \sum_{u \in D(v)} \eta^g_{v,u}$ where $A(v)$ is the set of ancestors of v and $D(v)$ descendants in the assembly graph
- The following is true $\phi^{g+}_v = \phi^{g-}_v = \phi^g_v$
- Strain intensities $\gamma_{g,s}$ as the rate per position that a read is generated from g in sample s
- Unitig lengths L_v
- Unitig bias θ_v is the fractional increase in reads generated from v given factors such as GC content influencing coverage
- Source node s and sink node t such that $\phi^{g+}_s = \phi^{g-}_s = \phi^{g+}_t = \phi^{g-}_t = 1$

Then assume normally distributed counts for each node in each sample giving a joint density for observations and latent variables:

$$P(\mathbf{X}, \mathbf{\Gamma}, \mathbf{H}, \mathbf{\Theta}) = \prod_{v=1}^{V} \prod_{s=1}^{S} \mathcal{N}(x_{v,s} | L_v \theta_v \sum_{h=1}^{G} \phi^h_v \gamma_{h,s}, \tau^{-1}) \prod_{h=1}^{G} \prod_{s=1}^{S} P(\gamma_{h,s}|\lambda_h)$$

$$\cdot \prod_{h=1}^{G} \prod_{v=1}^{V} \left[\phi^{h+}_v = \phi^{h-}_v\right] \left[\phi^{h-}_s = 1\right] \left[\phi^{h+}_t = 1\right] P(\tau)$$

$$\cdot \prod_{h=1}^{G} P(\lambda_h|\alpha_0, \beta_0) \prod_{v=1}^{V} P(\theta_v|\mu_0, \tau_0) \quad (2)$$

where [] indicates the Iverson bracket evaluating to 1 if the condition is true and zero otherwise. We assume an exponential prior for the $\gamma_{g,s}$ with a rate parameter that is strain dependent, such that:

$$P(\gamma_{g,s}|\lambda_g) = \lambda_g \exp(-\gamma_{g,s}\lambda_g) \quad (3)$$

We then place gamma hyper-priors on the λ_g:

$$P(\lambda_g|\alpha_0, \beta_0) = \frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)} \lambda_g^{\alpha_0 - 1} \exp(-\beta_0 \lambda_g) \quad (4)$$

This acts as a form of automatic relevance determination (ARD) forcing strains with low intensity across all samples to zero in every sample [8].
We use a standard Gamma prior for the precision:

$$P(\tau|\alpha_\tau, \beta_\tau) = \frac{\beta_\tau^{\alpha_\tau}}{\Gamma(\alpha_\tau)} \tau^{\alpha_\tau - 1} \exp(-\beta_\tau \tau) \quad (5)$$

For the biases θ_v we use a truncated normal prior:

$$\begin{aligned} P(\theta_v|\mu_0,\tau_0) = & \quad \frac{\sqrt{\frac{\tau_0}{2\pi}}\exp(-\frac{\tau_0}{2}(\theta_v-\mu_0)^2)}{1-\Psi(-\mu_0\sqrt{\tau_0})} \quad \theta_v >= 0 \\ = & \quad 0 \quad \theta_v < 0 \end{aligned}$$

where Ψ is the standard normal cumulative distribution. The mean of this is set to one, $\mu_0 = 1$, so that our prior is that the coverage on any given node is unbiased, with a fairly high precision $\tau_0 = 100$, to reflect an assumption that the observed coverage should reflect the summation of the strains. Finally, we assume a uniform prior over the possible discrete values of the $\eta^g_{v,u}$. If the assembly graph is a directed acyclic graph (DAG) then $\eta^g_{v,u} \in 0, 1$. We have found that for most genes and typical kmer lengths this is true, but we do not need to assume it.

Variational Approximation

We use variational inference to obtain an approximate solution to the posterior distribution of this model [7]. Variational inference is an alternative strategy to Markov chain Monte Carlo (MCMC) sampling. Rather than attempting to sample from the posterior distribution, variational inference assumes a tractable approximating distribution for the posterior, and then finds the parameters for that distribution that minimise the Kullback-Leibler divergence between the approximation and the true posterior distribution. Further, in mean-field variational inference the approximation can be factorised into a product over a number of components that each approximate the posterior of a parameter in the true distribution. In practice the Kullback-Leibler divergence is not computable because it depends on the evidence, i.e. the joint distribution marginalised over all latent variables. Instead, inference is carried out by maximising the evidence lower bound (ELBO), which is equal to the negative of the Kullback-Leibler divergence plus a constant, that constant being the evidence. In our case, because all the distributions are conjugate we can employ CAVI, coordinate ascent variational inference, to iteratively maximise the ELBO.

Our starting point is to assume the following factorisation for the variational approximation:

$$q(\mathbf{X},\mathbf{\Gamma},\mathbf{H}) = \prod_{h=1}^{G} q_h(\{\eta^h_{v,u}\}_{u,v\in\mathcal{A}}) \prod_{h=1}^{G}\prod_{s=1}^{S} q_h(\gamma_{h,s}) \prod_{h=1}^{G} q_h(\lambda_h) \prod_{v=1}^{V} q_v(\theta_v) q(\tau) \tag{6}$$

where $\mathcal{A}$ is the set of edges in the assembly graph and $\mathcal{V} = 1,\ldots,V$ the set of unitig sequence vertices. Note that we have assumed a fully factorised approximation except for the $\eta^h_{v,u}$, the paths for each strain through the graph. There we assume that the path for each strain forms a separate factor allowing strong correlations between the different elements of the path. This is therefore a form of structured variational inference [20].

To obtain the CAVI updates we use the standard procedure of finding the log of the optimal distributions q for each set of factorised variables as the expectation of the log joint distribution Equation 2 over all the other variables, except the one being optimised. Using an informal notation we will denote these expectations as $\langle \ln P \rangle_{-q_j}$ where q_j is the variable being optimised.

Then the mean field update for each set of $\{\eta^g_{v,u}\}_{u,v\in A}$ is derived as:

$$\begin{aligned}\ln q^*_g(\{\eta^g_{v,u}\}_{u,v\in A}) &= \langle \ln P\rangle_{-\eta^g_{v,u}u,v\in\mathcal{A}}\\ &= \ln\left(\prod_{v=1}^{V}\delta_{\phi_v^{g+},\phi_v^{g-}}\delta_{\phi_s^{g-},1}\delta_{\phi_t^{g+},1}\right)\\ &\quad-\left\langle\sum_{v=1}^{V}\sum_{s=1}^{S}\frac{\tau}{2}\left(x_{v,s}-\theta_v L_v[\sum_{h=1}^{G}\phi_v^h\gamma_{h,s}]\right)^2\right\rangle_{-\eta^g_{v,u}u,v\in\mathcal{A}}\\ &\quad+\text{Terms independent of }\eta^g\end{aligned}$$

Consider the second term only:

$$-\frac{\langle\tau\rangle}{2}\left(-\sum_{v=1}^{V}\sum_{s=1}^{S}2x_{v,s}L_v\langle\theta_v\rangle\langle\gamma_{g,s}\rangle\phi_v^g+L_v^2\langle{\theta_v}^2\rangle\langle(\sum_{h=1}^{G}\phi_v^h\gamma_{h,s})(\sum_{i=1}^{G}\phi_v^i\gamma_{i,s})\rangle\right)$$

This becomes:

$$-\frac{\langle\tau\rangle}{2}\left(\sum_{v=1}^{V}\sum_{s=1}^{S}\left[-2x_{v,s}\langle\theta_v\rangle L_v\langle\gamma_{g,s}\rangle\phi_v^g+2L_v^2\langle{\theta_v}^2\rangle\sum_{h\neq g}^{G}\langle\phi_v^h\rangle\langle\gamma_{h,s}\rangle\langle\gamma_{g,s}\rangle\phi_v^g+L_v^2\langle{\theta_v}^2\rangle\langle\gamma_{g,s}^2\rangle(\phi_v^g)^2\right]\right)$$

Which can be reorganised to:

$$\ln q^*_g(\{\eta^g_{v,u}\}_{u,v\in A})=\ln\left(\prod_{v=1}^{V}\delta_{\phi_v^{g+},\phi_v^{g-}}\delta_{\phi_s^{g-},1}\delta_{\phi_t^{g+},1}\right)+\sum_{v=1}^{V}c_{1,v}\phi_v^g+\sum_{v=1}^{V}c_{2,v}(\phi_v^g)^2 \tag{7}$$

Where:

$$\begin{aligned}c_{1,v} &= -\frac{\langle\tau\rangle}{2}\sum_{s=1}^{S}\left[-2x_{v,s}\langle\theta_v\rangle L_v\langle\gamma_{g,s}\rangle+2L_v^2\langle{\theta_v}^2\rangle\sum_{h\neq g}^{G}\langle\phi_v^h\rangle\langle\gamma_{h,s}\rangle\langle\gamma_{g,s}\rangle\right]\\ c_{2,v} &= -\frac{\langle\tau\rangle}{2}L_v^2\langle{\theta_v}^2\rangle\langle\gamma_{g,s}^2\rangle\end{aligned}$$

It is apparent from Equation 7 that the $q^*_g(\{\eta^g_{v,u}\}_{u,v\in A})$ takes the form of a multivariate discrete distribution with $|u,v\in A|$ dimensions. The first term in Equation 7 enforces the flow constraints, and does not separate across nodes, the next two terms are effectively coefficients on the total flow through a unitig and its square. The updates for the other variables below, depend on the expected values of the total flow through each of the unitig nodes for the strain g, $\langle\phi_v^g\rangle$, which themselves depend on the $\eta^g_{v,u}$. These expected values can be efficiently obtained for all v by representing Equation 7 as a factor graph comprising nodes consisting of factors corresponding to both the constraints and the flow probabilities through each node with variables $\eta^g_{v,u}$. We can then find the marginal probabilities for both the $\eta^g_{v,u}$ and the ϕ_v^g using the Junction Tree algorithm [41], from these we can calculate the required expectations. Next we consider the mean field update for the $\gamma_{g,s}$:

$$\begin{aligned}\ln q^*(\gamma_{g,s}) &= \langle\ln P\rangle_{-\gamma_{g,s}}\\ &= -\left\langle\sum_{v=1}^{V}\frac{\tau}{2}\left(x_{v,s}-\theta_v L_v[\sum_{h=1}^{G}\phi_v^h\gamma_{h,s}]\right)^2\right\rangle_{-\gamma_{g,s}}-\langle\lambda_g\rangle\gamma_{g,s}\end{aligned}$$

$$\begin{aligned}
&\ln q^*(\gamma_{g,s}) = \\
&\quad -\frac{\langle\tau\rangle}{2}\left(\sum_{v=1}^{V} -2x_{v,s}\langle\theta_v\rangle L_v\langle\phi_v^g\rangle\gamma_{g,s} + 2\langle\theta_v^2\rangle L_v^2\gamma_{g,s}\langle\phi_v^g\rangle\sum_{h\neq g}\langle\gamma_{h,s}\rangle\langle\phi_v^h\rangle + \langle\theta_v^2\rangle L_v^2\gamma_{g,s}^2\langle[\phi_v^g]^2\rangle\right) \\
&\quad -\langle\lambda_g\rangle\gamma_{g,s}
\end{aligned}$$

with the restriction $\gamma_{g,s} > 0$ this gives a normal distribution but truncated to $(0, \inf)$ for $\gamma_{g,s}$, with mean and precision:

$$\mu_{g,s} = \frac{\sum_v x_{v,s}\theta_v L_v\langle\phi_v^g\rangle - \langle\phi_v^g\rangle\sum_{h\neq g}\langle\gamma_{h,s}\rangle\langle\phi_v^h\rangle\langle\theta_v^2\rangle L_v^2}{\sum_v L_v^2\langle[\phi_v^g]^2\rangle} - \frac{\langle\lambda_g\rangle}{\tau_{g,s}} \tag{8}$$

$$\tau_{g,s} = \langle\tau\rangle\sum_v L_v^2\langle[\phi_v^g]^2\rangle \tag{9}$$

$$\tag{10}$$

Derivations for the other updates follow similarly giving a Gamma posterior for the τ with parameter updates:

$$\alpha_\tau = \alpha_0 + \Omega/2 \tag{11}$$

$$\beta_\tau = \beta_0 + \sum_{v,s}\langle(x_{v,s} - \lambda_{v,s})^2\rangle \tag{12}$$

where $\Omega = VS$ and we have used $\lambda_{v,s}$ as a short hand for the predicted count number:

$$\lambda_{v,s} = \theta_v\sum_g \gamma_{g,s}\phi_v^g.$$

Then the τ have the following expectations and log expectations:

$$\langle\tau_{v,s}\rangle = \alpha_\tau/\beta_\tau \tag{13}$$

$$\langle\log\tau_{v,s}\rangle = \psi\left(\alpha_\tau + 1/2\right) - \log\left(\beta_\tau\right) \tag{14}$$

where ψ is the digamma function. The biases θ_v have a truncated normal distribution and their updates can be derived similar to the above.

Evidence lower bound (ELBO)

Iterating the CAVI updates defined above will generate a variational approximation that is optimal in the sense of maximising the evidence lower bound (ELBO) so called because it bounds the log evidence, $log(p(x)) \geq ELBO(q(z))$. It is useful to calculate the ELBO whist performing CAVI updates to verify convergence and the ELBO itself is sometimes used as a Bayesian measure of model fit, although as a bound that may be controversial [7]. The ELBO can be calculated from the relationship:

$$ELBO(q) = \mathbb{E}\left[\log p(x|z)\right] + \mathbb{E}\left[\log p(z)\right] - \mathbb{E}\left[\log q(z)\right] \tag{15}$$

The first term is simply the expected log-likelihood of the data given the latent variables. In our case it is:

$$\mathbb{E}\left[\log p(x|z)\right] = \frac{1}{2}\Omega\left(\langle\log\tau\rangle - \log(2\pi)\right) - \frac{1}{2}\langle\tau\rangle\langle(x_{v,s} - L_v\theta_v\sum_{h=1}^{G}\phi_v^h\gamma_{h,s})^2\rangle \tag{16}$$

where $\Omega = VS$ and the expectations are over the optimised distributions q.

The second term is the expectation under $q(z)$ of the log prior distributions. In our case with standard distributions it is easy to calculate for instance for each of the $\gamma_{g,s}$:

$$\mathbb{E}\left[\log p(\gamma_{g,s})\right] = \frac{1}{2}\log\left(\frac{\tau_{g,s}}{2\pi}\right) - \frac{\tau_{g,s}}{2}\left(\langle\gamma_{g,s}^2\rangle + \mu_{g,s}^2 - 2\mu_{g,s}\langle\gamma_{g,s}\rangle\right) - \log\left[\frac{1}{2}\operatorname{erf}\left(\mu_{g,s}\sqrt{\frac{\tau_{g,s}}{2}}\right)\right].$$

With the $\mu_{g,s}$ and $\tau_{g,s}$ given by their current values derived from Equation 10 and the moments of $\gamma_{g,s}$ calculated from a truncated normal distribution with those current parameters. The third terms are simply the negative entropy of the variational approximations and for the standard distributions used here are easily calculated.

Implementation details

One update of the algorithm consists of updating each variable or sets of variables in turn given the current optimal solutions of the other distributions. In practice we update:

- Compute the marginal flows $\{\eta_{v,u}^g\}_{u,v\in A}$ for each strain $g = 1,\ldots,G$ in turn using Equation 7 and the Junction Tree algorithm. This can be performed for each single copy-core gene independently
- Update the truncated normal strain abundances $q(\gamma_{g,s})$ for each strain in each sample, $s = 1,\ldots,S$ using Equation 10
- Update the $q(\tau)$
- Update the ARD parameter distributions $q(\lambda_g)$ if used
- Update the nodes biases $q(\theta_v)$
- Check for redundant or low abundance strains and remove (see below)

After a maximum fixed number of iterations or if the ELBO converges we stop iterating. Variational inference can be sensitive to initial conditions as it can only find local maxima of the ELBO, we therefore use a previously published variational Bayesian version of non-negative matrix factorisation [8], to find an initial approximate solution.

Empirical modelling of node variances

For low-coverage MAGs a precision that is identical for all nodes performs satisfactorily, but since the true distribution of read counts is Poisson this overestimates the precision for nodes with high counts $x_{v,s}$. To address we developed an empirical procedure where we first calculate $\langle\log\tau_{v,s}\rangle$ for each node using Equation 14 as:

$$\langle\log\tau_{v,s}\rangle = \psi\left(\alpha_0 + 1/2\right) - \log\left(\beta_0 X_{v,s} + \frac{1}{2}\langle(x_{v,s} - \lambda_{v,s})^2\rangle\right) \tag{17}$$

a quantity which exhibits high variability, so we then smooth this over $\log(x_{v,s})$ using generalised additive models as implemented in pyGAM [36] to give $\langle \log \tau_{v,s} \rangle^*$. The term $\beta_0 X_{v,s}$ gives a prior which is effectively Poisson. We then obtain $\langle \tau_{v,s} \rangle$ as $\exp(\langle \log \tau_{v,s} \rangle^*)$. This procedure has no theoretical justification but gives good results in practice. This approach of modelling a non-Gaussian distribution as a Gaussian with empirical variances is similar to that used in voom for RNASeq [23].

Cross-validation to determine optimum number of strains

The ARD procedure usually converges on the correct number of strains except for high-coverage MAGs where overfitting may occur and too many strains can be predicted. We therefore additionally implemented a cross-validation procedure, splitting the entire data matrix $x_{v,s}$ into test and train folds (default ten folds) and training the model on the train fold and then testing on the held out data. The correct number of strains was then taken to be the one that maximised the log predictive posterior with an empirical variance reflecting the Poisson nature of the data. The exact test statistic being:

$$\sum_{v,s \in E} \frac{1}{2} \log(\tau'_{v,s}) - \frac{1}{2} \sum_{v,s \in E} \tau'_{v,s} \langle (x_{v,s} - \lambda_{v,s})^2 \rangle \qquad (18)$$

where $\tau'_{v,s} = 1/(0.5 + x_{v,s})$ and E indicates data points in the test set to down-weight high read count nodes reflecting approximately Poisson noise.

Nanopore sequence analysis

Sequence preprocessing

To enable a qualitative comparison between haplotypes obtained from the Nanopore reads and the BayesPaths predictions we developed the following pipeline applied at the level of individual single-copy core genes (SCGs) from MAGs:

1. We mapped all reads using minimap2 [26] against the SCG contig consensus ORF sequence and selected those reads with alignments that spanned at least one third of the gene with a nucleotide identity $> 80\%$.

2. We then extracted the aligned portion of each read, reverse complementing if necessary, so that all reads ran in the same direction as the SCG ORF.

3. We then obtained the variant positions on the consensus from the output of the DESMAN pipeline [32]. These are variant positions prior to haplotype calling representing the total unlinked variation observed in the short reads.

4. For each Nanopore fragment we aligned against the SCG ORF using a Needleman-Wunsch global alignment and generated a condensed read comprising bases only from the short read variant positions.

This provided us with a reduced representation of each Nanopore read effectively filtering variation that was not observed in the short reads. These reduced representations were then used to calculate distances, defined as Hamming distances on the variant positions normalised by number of positions observed, both between the reads and between the reads and the predicted COG sequences from BayesPaths. From these we generated NMDS plots indicating sequence diversity, and they provided an input to the hybrid Nanopore strain resolution algorithm below.

EM algorithm for hybrid Nanopore strain resolution

We also developed a simple EM algorithm for direct prediction of paths and their abundances on the short read assembly graph that are consistent with a set of observed long reads. We began by mapping the set of $n = 1, \ldots, N$ Nanopore sequences denoted $\{\mathcal{S}_n\}$ onto the corresponding simplified SCG graph generated by STRONG using GraphAligner [33]. This provided us with N optimal alignment paths as walks in our SCG graph. We denote this graph $\mathcal{G}$ comprising unitig vertices v and edges $e \in \{u, v\}$ defining overlaps.

We assume, as is almost always the case that the graphs contain no unitigs in both forward and reverse configurations, and that there are no cycles, so that each SCG is a directed acyclic graph (DAG) with one copy of each unitig, and we only need to track the direction that each overlap enters and leaves each unitig. Then best alignment walks comprise a sequence of edges, $e^n_1, \ldots, e^n_{W_n}$ where W_n is the number of edges in the walk of read n, that traverse the graph.

Given these observed Nanopore reads we aim to reconstruct G haplotypes comprising paths from a dummy source node s, which has outgoing edges to all the true source nodes in the graph, through the graph to a dummy sink t, which connects all the true sinks. We further assume that each haplotype has relative frequency π_g. Each such haplotype path $\mathcal{P}_g = \{s, e^g_1, \ldots, e^g_{W_g}, t\}$ will translate into a nucleotide sequence $\mathcal{T}_g$. We assume that these haplotypes generate Nanopore reads with a fixed error rate ϵ which gives a likelihood:

$$P(\{\mathcal{S}_1, \ldots, \mathcal{S}_N\}|\pi, \{\mathcal{T}_1, \ldots, \mathcal{T}_G\}) = \prod_{n=1}^{N} \left(\sum_{g=1}^{G} \pi_g \epsilon^{m_{n,g}} (1-\epsilon)^{M_{n,g}} \right). \tag{19}$$

where $m_{n,g}$ is the number of basepair mismatches between $\mathcal{S}_n$ and $\mathcal{T}_g$ counting insertions, deletions and substitutions equally and $M_{n,g}$ the number of matches.

To maximise this likelihood we used an Expectation-Maximisation algorithm. Iterating the following steps until convergence:

1. E-step: Calculate the responsibility of each haplotype for each sequence as:

$$z_{n,g} = \frac{\pi_g \epsilon^{m_{n,g}} (1-\epsilon)^{M_{n,g}}}{\sum_h \pi_h \epsilon^{m_{n,h}} (1-\epsilon)^{M_{n,h}}} \tag{20}$$

   Alignments of reads against haplotypes were performed using vsearch [34].

2. M-step: We update each haplotype by finding the most likely path on the short read graph given the current expectations. These are calculated by assigning a weight w^g_e to each edge e in the graph as:

$$w^g_e = \sum_{n \in e} z_{n,g} L_e \tag{21}$$

   where $n \in e$ are the set of reads whose optimal alignment contains that edge and L_e is the unique length of the unitig the edge connects to, *i.e.* ignoring the overlap length. We then find for haplotype g the maximal weight path through this DAG using a topological sort. The error rates are updated as:

$$\epsilon = \frac{\sum_n \sum_g z_{n,g} m_{n,g}}{\sum_n \sum_g z_{n,g} L_{n,g}} \tag{22}$$

where $L_{n,g}$ are the alignment lengths.

As is often the case with EM algorithms convergence depends strongly on initial conditions. Therefore we initialise using a partitioning around medoids clustering using the distances calculated in Methods - Nanopore Sequence Analysis. We can estimate the number of haplotypes from the negative log-likelihood as a function of haplotype number.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

The anaerobic digester time series have been uploaded to the ENA as part of the study PRJEB39861. The STRONG pipeline and synthetic community data are available from `https://github.com/chrisquince/STRONG` and the BayesPaths algorithm `https://github.com/chrisquince/BayesPaths`. The code for synthetic community generation from `https://github.com/chrisquince/STRONG_Sim` and the Nanopore EM algorithm `https://github.com/chrisquince/NanoHap`.

Competing interests

The authors declare that they have no competing interests.

Funding

This work for made possible through the MRC Methodology Grant ‘Strain resolved metagenomics for medical microbiology’ MR/S037195/1. CQ is also funded through MRC fellowship (MR/M50161X/1) as part of the CLoud Infrastructure for Microbial Genomics (CLIMB) consortium (MR/L015080/1). SR is funded through BBSRC ‘EBI Metagenomics - enabling the reconstruction of microbial populations’ (BB/R015171/1). OSS acknowledges funding through the BBSRC (BB/N023285/1 and BB/L502029/1).

Authors' contributions

CQ devised and coded the BayesPaths algorithm, assisted with the creation of the STRONG pipeline, analysed results and wrote the MS. SN devised and coded the graph algorithms, assisted with the creation of the STRONG pipeline, and edited the MS. SR coded the STRONG pipeline. RJ generated and processed the AD Nanopore sequence data. OSS helped devise the AD sequencing study. JKS assisted with the STRONG pipeline and edited the MS. AL contributed to the creation of the algorithms and edited the MS. AME contributed to the creation of the algorithms and assisted with figures. RC tested the STRONG pipeline, contributed to the creation of the algorithms and edited the MS. AED helped plan the STRONG pipeline, assisted the creation of the algorithms and edited the MS.

Acknowledgements

The subgraph extraction procedure was developed following discussions with Tatiana Dvorkina (PhD student in SPbSU).

References

1. T.-H. Ahn, J. Chai, and C. Pan. Sigma: Strain-level inference of genomes from metagenomic analysis for biosurveillance. *Bioinformatics*, 31(2):170–177, 09 2014.

2. M. Albertsen, P. Hugenholtz, A. Skarshewski, K. L. Nielsen, G. W. Tyson, and P. H. Nielsen. Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nature Biotechnology*, 31:533 EP –, May 2013.

3. J. Alneberg, B. S. Bjarnason, I. de Bruijn, M. Schirmer, J. Quick, U. Z. Ijaz, L. Lahti, N. J. Loman, A. F. Andersson, and C. Quince. Binning metagenomic contigs by coverage and composition. *Nature Methods*, 11(11):1144–1146, sep 2014.

4. J. A. Baaijens, B. Van der Roest, J. Köster, L. Stougie, and A. Schönhuth. Full-length de novo viral quasispecies assembly through variation graph construction. *Bioinformatics*, 05 2019. btz443.

5. E. Bernard, L. Jacob, J. Mairal, and J.-P. Vert. Efficient RNA isoform identification and quantification from RNA-Seq data with network flows. *Bioinformatics*, 30(17):2447–2455, 05 2014.

6. E. Bernard, L. Jacob, J. Mairal, E. Viara, and J.-P. Vert. A convex formulation for joint RNA isoform detection and quantification from multiple RNA-seq samples. *BMC Bioinformatics*, 16(1):262, 2015.

7. D. M. Blei, A. Kucukelbir, and J. D. McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017.

8. T. Brouwer, J. Frellsen, and P. Lió. Comparative study of inference methods for bayesian nonnegative matrix factorisation. In M. Ceci, J. Hollmén, L. Todorovski, C. Vens, and S. Džeroski, editors, *Machine Learning and Knowledge Discovery in Databases*, pages 513–529, Cham, 2017. Springer International Publishing.

9. C. T. Brown, L. A. Hug, B. C. Thomas, I. Sharon, C. J. Castelle, A. Singh, M. J. Wilkins, K. C. Wrighton, K. H. Williams, and J. F. Banfield. Unusual biology across a group comprising more than 15% of domain bacteria. *Nature*, 523:208 EP –, Jun 2015.

10. C. T. Brown, D. Moritz, M. P. O'Brien, F. Reidl, T. Reiter, and B. D. Sullivan. Exploring neighborhoods in large metagenome assembly graphs using spacegraphcats reveals hidden sequence diversity. *Genome Biology*, 21(1):164, 2020.

11. P.-A. Chaumeil, A. J. Mussig, P. Hugenholtz, and D. H. Parks. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics*, 36(6):1925–1927, 11 2019.

12. P. I. Costea, G. Zeller, S. Sunagawa, E. Pelletier, A. Alberti, F. Levenez, M. Tramontano, M. Driessen, R. Hercog, F.-E. Jung, J. R. Kultima, M. R. Hayward, L. P. Coelho, E. Allen-Vercoe, L. Bertrand, M. Blaut, J. R. M. Brown, T. Carton, S. Cools-Portier, M. Daigneault, M. Derrien, A. Druesne, W. M. de Vos, B. B. Finlay, H. J. Flint, F. Guarner, M. Hattori, H. Heilig, R. A. Luna, J. van Hylckama Vlieg, J. Junick, I. Klymiuk, P. Langella, E. Le Chatelier, V. Mai, C. Manichanh, J. C. Martin, C. Mery, H. Morita, P. W. O'Toole, C. Orvain, K. R. Patil, J. Penders, S. Persson, N. Pons, M. Popova, A. Salonen, D. Saulnier, K. P. Scott, B. Singh, K. Slezak, P. Veiga, J. Versalovic, L. Zhao, E. G. Zoetendal, S. D. Ehrlich, J. Dore, and P. Bork. Towards standards for human fecal sample processing in metagenomic studies. *Nature Biotechnology*, 35(11):1069–1076, 2017.

13. K. DD, F. J, E. R, and W. Z. Metabat, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ*, 3:e1165, 2015.

14. T. O. Delmont, C. Quince, A. Shaiber, Ö. C. Esen, S. T. M. Lee, M. S. Rappé, S. L. McLellan, S. Lücker, and A. M. Eren. Nitrogen-fixing populations of Planctomycetes and Proteobacteria are abundant in surface ocean metagenomes. *Nature Microbiology*, 3(7):804–813, 2018.

15. L. Dijkshoorn, B. M. Ã. Ursing, and J. B. Ursing. Strain , clone and species : comments on three basic concepts of bacteriology. *J. Med. Microbiol.*, 49:397–401, 2000.

16. A. Eren, O. Esen, C. Quince, J. Vineis, H. Morrison, M. Sogin, and T. Delmont. Anvi'o: an advanced analysis and visualization platform for 'omics data. *PeerJ*, 3:e1319, 2015.

17. L. E.S. and W. M.S. Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics*, 2:231–239, 1988.

18. E. Garrison and G. Marth. Haplotype-based variant detection from short-read sequencing. *arXiv e-prints*, page arXiv:1207.3907, July 2012.

19. E. Garrison, J. Sirén, A. M. Novak, G. Hickey, J. M. Eizenga, E. T. Dawson, W. Jones, S. Garg, C. Markello, M. F. Lin, B. Paten, and R. Durbin. Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nature Biotechnology*, 36(9):875–879, 2018.

20. M. Hoffman and D. Blei. Stochastic Structured Variational Inference. In G. Lebanon and S. V. N. Vishwanathan, editors, *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, volume 38 of *Proceedings of Machine Learning Research*, pages 361–369, San Diego, California, USA, 09–12 May 2015. PMLR.

21. W. Huang, L. Li, J. R. Myers, and G. T. Marth. ART: a next-generation sequencing read simulator. *Bioinformatics*, 28(4):593–594, 12 2011.

22. J. Köster and S. Rahmann. Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics*, 28(19):2520–2522, 08 2012.

23. C. W. Law, Y. Chen, W. Shi, and G. K. Smyth. voom: precision weights unlock linear model analysis tools for rna-seq read counts. *Genome Biol.*, 15:R29, 2014.

24. A. Leimbach, J. Hacker, and U. Dobrindt. *E. coli as an All-Rounder: The Thin Line Between Commensalism and Pathogenicity*, pages 3–32. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013.

25. D. Li, C.-M. Liu, R. Luo, K. Sadakane, and T.-W. Lam. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics*, 31(10):1674–1676, 01 2015.

26. H. Li. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, 34(18):3094–3100, 05 2018.

27. C. Luo, R. Knight, H. Siljander, M. Knip, R. J. Xavier, and D. Gevers. Constrains identifies microbial strains in metagenomic datasets. *Nature Biotechnology*, 33:1045 EP –, Sep 2015.

28. S. Nurk, D. Meleshko, A. Korobeynikov, and P. Pevzner. metaspades: a new versatile metagenomic assembler. *Genome Res.*, 27:824–834, 2017.

29. J. D. O'Brien, X. Didelot, Z. Iqbal, L. Amenga-Etego, B. Ahiska, and D. Falush. A bayesian approach to inferring the phylogenetic structure of communities from metagenomic data. *Genetics*, 197(3):925–937, 2014.

30. M. G. Pachiadaki, J. M. Brown, J. Brown, O. Bezuidt, P. M. Berube, S. J. Biller, N. J. Poulton, M. D. Burkart, J. J. La Clair, S. W. Chisholm, and R. Stepanauskas. Charting the complexity of the marine microbiome through single-cell genomics. *Cell*, 179(7):1623 – 1635.e11, 2019.

31. E. Pasolli, F. Asnicar, S. Manara, M. Zolfo, N. Karcher, F. Armanini, F. Beghini, P. Manghi, A. Tett, P. Ghensi, M. C. Collado, B. L. Rice, C. DuLong, X. C. Morgan, C. D. Golden, C. Quince, C. Huttenhower, and N. Segata. Extensive unexplored human microbiome diversity revealed by over 150,000 genomes from metagenomes spanning age, geography, and lifestyle. *Cell*, 176(3):649–662.e20, Jan 2019.

32. C. Quince, T. O. Delmont, S. Raguideau, J. Alneberg, A. E. Darling, G. Collins, and A. M. Eren. Desman: a new tool for de novo extraction of strains from metagenomes. *Genome Biology*, 18(1):181, 2017.

33. M. Rautiainen, V. Mäkinen, and T. Marschall. Bit-parallel sequence-to-graph alignment. *Bioinformatics*, 35(19):3599–3607, 03 2019.

34. T. Rognes, T. Flouri, B. Nichols, C. Quince, and F. Mahé. Vsearch: a versatile open source tool for metagenomics. *PeerJ*, 4:e2584, 2016.

35. N. Segata. On the road to strain-resolved comparative metagenomics. *mSystems*, 3(2), 2018.

36. D. Servén and C. Brummitt. pygam: Generalized additive models in python. *Zenodo*, 2018.

37. R. L. Tatusov, M. Y. Galperin, D. A. Natale, and E. V. Koonin. The COG database : a tool for genome-scale analysis of protein functions and evolution. *Nucl. Acid Res.*, 28(1):33–36, 2000.

38. D. T. Truong, A. Tett, E. Pasolli, C. Huttenhower, and N. Segata. Microbial strain-level population structure and genetic diversity from metagenomes. pages 626–638, 2017.

39. T. Van Rossum, P. Ferretti, O. M. Maistrenko, and P. Bork. Diversity within species: interpreting strains in microbiomes. *Nature Reviews Microbiology*, 2020.

40. M. Vos and X. Didelot. A comparison of homologous recombination rates in bacteria and archaea. *The ISME Journal*, 3(2):199–208, 2009.

41. M. J. Wainwright and M. I. Jordan. Graphical models, exponential families, and variational inference. *Found. Trends Mach. Learn.*, 1(1-2):1–305, Jan. 2008.

42. Z. Zhou, N. Luhmann, N.-F. Alikhan, C. Quince, and M. Achtman. Accurate reconstruction of microbial strains from metagenomic sequencing using representative reference genomes. In B. J. Raphael, editor, *Research in Computational Molecular Biology*, pages 225–240, Cham, 2018. Springer International Publishing.