doi: DOI HERE
Paper

PAPER

Reference-free Structural Variant Detection in Microbiomes via Long-read Coassembly Graphs

Kristen D. Curry,[1,2,*] Feiqiao Brian Yu,[3] Summer E. Vance,[4] Santiago Segarra,[5] Devaki Bhaya,[6] Rayan Chikhi,[7] Eduardo P.C. Rocha[2] and Todd J. Treangen[1]

[1]Rice University, Department of Computer Science, Houston, TX 77005, United States, [2]Institut Pasteur, Université Paris Cité, CNRS, UMR3525, Microbial Evolutionary Genomics, 75015 Paris, France, [3]Arc Institute, Palo Alto, CA 94304, United States, [4]University of California, Berkeley, Department of Environmental Science, Policy, and Management, Berkeley, CA 94720, United States, [5]Rice University, Department of Electrical and Computer Engineering, Houston, TX 77005, United States, [6]Carnegie Institution for Science, Department of Plant Biology, Stanford, CA 94305, United States and [7]Institut Pasteur, Université Paris Cité, Sequence Bioinformatics unit, 75015 Paris, France

[*]Corresponding author. kristen.d.curry@rice.edu

Abstract

Bacterial genome dynamics are vital for understanding the mechanisms underlying microbial adaptation, growth, and their broader impact on host phenotype. Structural variants (SVs), genomic alterations of 10 base pairs or more, play a pivotal role in driving evolutionary processes and maintaining genomic heterogeneity within bacterial populations. While SV detection in isolate genomes is relatively straightforward, metagenomes present broader challenges due to absence of clear reference genomes and presence of mixed strains. In response, our proposed method rhea, forgoes reference genomes and metagenome-assembled genomes (MAGs) by encompassing a single metagenome coassembly graph constructed from all samples in a series. The log fold change in graph coverage between subsequent samples is then calculated to call SVs that are thriving or declining throughout the series. We show rhea to outperform existing methods for SV and horizontal gene transfer (HGT) detection in two simulated mock metagenomes, which is particularly noticeable as the simulated reads diverge from reference genomes and an increase in strain diversity is incorporated. We additionally demonstrate use cases for rhea on series metagenomic data of environmental and fermented food microbiomes to detect specific sequence alterations between subsequent time and temperature samples, suggesting host advantage. Our innovative approach leverages raw read patterns rather than references or MAGs to include all sequencing reads in analysis, and thus provide versatility in studying SVs across diverse and poorly characterized microbial communities for more comprehensive insights into microbial genome dynamics.

Key words: metagenome, microbiome, structural variants, gene transfer, long-read sequencing

Introduction

Structural variants (SVs), loosely defined as genomic alterations that are 10 base pairs (bps) or longer (12), play an important role in driving both evolutionary adaptation and heterogeneity in bacterial genomes (31). Bacterial genome dynamics not only influence the ability for the bacteria to grow and adapt to changing environments (32), but can also impact the function of the microbial community as a whole and the phenotype of the host (11). In isolate genomics, the goal of SV detection is relatively straightforward: detect long genomic differences between a sequence and reference genome that can be classified as an insertion, deletion, inversion, duplication, translocation, or any combination of the prior (37). However, in metagenomics, when reference genomes may not be well-defined and a mixed population of similar strains may exist in the community, detection of SVs becomes more complex (37).

SV detection methods can be broadly categorized into three groups: mapping-driven, assembly-driven, and pattern-driven (Table 1) (37). In mapping-driven approaches, reads are directly aligned to an established reference genomes or pangenome of sequences, then unexpected mapping patterns identify SVs. In assembly-driven approaches, reads are first assembled into longer sequences (contigs), then aligned to another contig or reference to detect long scale differences. In pattern-driven approaches, SV patterns are pre-defined then search for in sequencing reads. Zeevi et al. developed a mapping-driven SV detection approach for metagenomic short reads to survey SVs associated with host disease risk factors in the human gut microbiome (42). The authors built a comprehensive database specifically for known microbes in the human gut microbiome and developed an "iterative coverage-based read assignment" (ICRA) algorithm to repeatedly adjust read assignments and establish alignments. Their SGV-Finder algorithm then scans the coverage of each reference genome for presence of regions with unexpectedly low (deletions) or high (duplications) coverage. While this method has been effective as a comprehensive search for SVs in the human gut microbiome correlating to expressed phenotypes (24), relying on a confident database of reference genomes is challenging for communities that have not been extensively characterized. This pipeline is additionally restricted to only deletions and duplications relative to reference genomes in the supplied database.

Table 1. Methods for SV detection in metagenomes, separated by types: mapping(M)-, assembly(A)-, and pattern(P)-driven. SV types abbreviations are as follows: Ins: insertion, Del: deletion, Dup: duplication, Inv: inversion, Trans: translocation, and CI: complex indel (defined here as an insertion and deletion at the same location). Input types are short reads (short), long reads (long), or metagenome-assembled genomes (MAG).

Software	Type	Detected SV Types	Input
SVGFinder (42)	M	Ins, Dup	short
MetaSVs (23)	A	Ins, Del, Dup, Inv, Trans	MAG
MetaCHIP (35)	A	HGT Ins	MAG
PhaseFinder (16)	P	Inv	short
DIVE (1)	P	MGE Ins, MGE CI	short
Rhea	P	Ins, Del, Dup, CI	long

To expand upon the types of SVs detected and leverage advantages of long read technologies, MetaSVs, an assembly-driven approach, was designed (23). In this pipeline, long and short reads combined help to confidently create and classify metagenome-assembled genomes (MAGs). Each MAG is then evaluated independently through whole-genome alignment to a reference MAG or genome with the SV detection tool MUM & Co (29). Chen et al. utilized MetaSVs to expand upon characterized SVs in the human gut (notably insertions and inversions) and demonstrates the value in incorporating long reads for SV detection (9). However, this assembly-driven method is still highly dependent on a reference database, as it is the taxonomic reference-driven classifications that determine which MAGs get compared to which references. Additionally, unique MAGs are often not created for subtle SV differences (18), especially in microbial communities where similar strains are present (14).

MetaCHIP is another MAG-based approach for the slightly different goal of detecting recent horizontal gene transfer (HGT) events within a metagenome (36). In an HGT event, genetic material is exchanged between organisms (28), resulting in an insertion SV for the recipient microbe. MetaCHIP effectively evaluates each MAG in the community for a gene sequence that has more BLASTN (2) hits to genes in a different MAG than its own. This algorithm, however, can only detect insertion genes that are highly similar to another MAG, which resulted in simulation results declining at 25% mutation rate between donor and recipient.

To entirely avoid reference genomes and MAG creation, two pattern-driven methods have been developed. PhaseFinder (16) was created for detection of inversions in bacterial genomes from genomic or metagenomic data, by detecting regions flanked by inverted repeats where sequencing reads support both orientations. DIVE (1) was developed in 2023 to identify sequences surrounding genetic diversification such as transposable elements, within MGE variability hotspots, or CRISPR repeats, by detecting constant k-mers with diverse flanking sequences to define MGE bounding sequences and transposon arms. While both these methods show how patterns in raw read can be used to eliminate reference genomes and MAGs, they are limited to only these specific patterns.

Rhea takes a different approach to detect SV patterns within a microbial community. It constructs a coassembly graph from all metagenomes in a series that are expected to have similar communities (i.e. longitudinal time series or cross-sectional studies where a significant portion of the strains are shared across samples). Regions of the graph indicative of SVs are then highlighted, as previously explored for characterization of genome variants (27; 13). The log fold change in graph coverage between consecutive steps in the series is then used to reduce false SV calls made from assembly error, account for shifting levels of microbe relative abundance, and ultimately permit SV detection in understudied and complex microbial environments. Recent work utilizes coassembly graphs for metagenomes to decompose strain diversity into haplotypes (30), but to the best of our knowledge, this is the first time coassembly graph patterns have been used for automated detection of SVs in a metagenome series.

Methods

Rhea method

Rhea takes as input a series of long-read metagenomic sequences, expected to be taken from the same source at different time points or some other step-wise metadata separation. A single metagenome assembly graph is constructed by combining all provided samples, then each sample is separately aligned back to the graph. Change in graph coverage between subsequent samples and the graph structure are used to call SVs (Figure 1).

SV definitions

Four types of SVs are detected in rhea: insertions, deletions, tandem duplications (37), and complex indels (41; 33). An insertion here is a sequence that has been integrated in increasing abundance between subsequent steps in the series. A deletion is the opposite, a subsequence that is declining. A tandem duplication is a gene sequence that has been repeated, directly one after another, in increasing presence. A complex indel as a sequence that has drastically changed between subsequent steps, showing the signature of a deletion and insertion at the same location. In this pipeline, SV detection equates to an increase in abundance of the SV, rather than simply a novel appearance, and therefore suggests a provided advantage for the host microbe or the community.

Graph construction and coverage calculations

A single coassembly graph for the series with N samples is constructed by combining all reads from all samples into one metaFlye run (19), with `--keep-haplotypes` parameter set to true to maintain strain variations. After the graph is constructed, each sample is separately aligned back to the graph with minigraph (22). An undirected graph is then built mimicking the structure of the metaFlye assembly graph where a single node is drawn for each complementary pair, as seen in the assembly graph visualization software Bandage "single" option (38). This graph is defined as $G = (V, E)$ with a set of k nodes $V = \{v_1, v_2.., v_k\}$ and a set of edges E. Each edge $(e_{i,j})$ is then given a weight equal to the number of edges that appear between nodes i and j in the metaFlye assembly graph, given

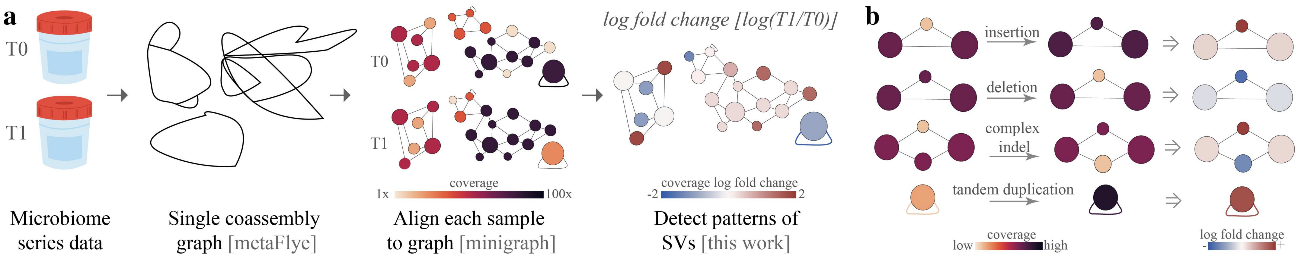


Fig. 1. (a) To utilize rhea, first, microbiome series data must be collected and long whole genome sequencing reads generated. Then, within rhea, a coassembly graph of all reads in the series is created with metaFlye. Reads from each sample are then separately aligned to the coassembly graph with minigraph. Rhea evaluates log fold change in coverage between series steps for SV-specific patterns in the assembly graph to detect structural variants between steps. (b) Assembly graph patterns detected in rhea, which indicate potential insertions, deletions, complex indels, and tandem duplicates. Insertions and deletions are detected by observing a triangle where one node has a significantly higher (insertion) or lower (deletion) log fold change. Complex indels are noted by a square with one or two outliers; in the case of two outliers, the two outliers must be of opposing sides of the median and not have an edge between them. Tandem duplicates are detected by a log fold change of a self-loop edge coverage greater than 1.

there exist at least one edge between i and j in the assembly graph. Each edge ($e_{i,j}$) thus denotes the existence of overlap reads that expand directly from v_i to v_j (or from v_j to v_i) without gaps, in either direction (forward or reverse) for the sequences in i and j. Minigraph alignments are then used to calculate node and edge coverage for each step in the series. Node coverage is calculated as the average coverage per base pair within the node, calculated by summing the coverage for each base pair divided by the total number of base pairs in the node. To account for error, all nodes with coverage less than 1, are set to a coverage of 1. Node coverage is then normalized for the entire series, by first calculating the median total base pairs m across samples in the series, then establishing a multiplier for each sample $n = 0..N$ as bp_n/m, where bp_n is the number of base pairs in sample n. This multiplier for each step is applied to all node coverage for each $n = 0..N$. Edge coverage for each edge $e_{i,j}$ at each step n in the series is counted as the number of occurrence a read path covers directly from i to j or j to i in the read-graph alignment for step n. Each node in our undirected assembly graph then holds a vector of log fold change in coverage between subsequent steps in the series, calculated for each node i as $log(vc_{i,t_n}/vc_{i,t_{n-1}})$, where vc_{i,t_n} is the coverage of node i at step n in the series for all steps $n = 1...N$. A log fold change vector is also assigned to each edge (i,j), defined as $log(ec_{(i,j),t_n}/ec_{(i,j),t_{n-1}})$, where ec_{i,t_n} is the coverage of edge $e_{i,j}$ at step n in the series for all steps $n = 1..N$. The log fold change vectors are then used in the next step to detect SVs and account for assembly error and changes in genome relative abundance between subsequent samples.

Detected SV graph patterns

Rhea utilizes the graph structure, edge weights, and the log fold change coverage vectors to call SVs between each pair of consecutive samples in the series. For insertions and deletions, each triangle is searched for the pattern of two similar log fold change values and one that is significantly different for each step. This is completed by: calculating the median and standard deviation between the three log fold change, then labeling any node with a value that is more than one standard deviation away from the median as an outlier. If the triangle contains exactly one outlier, then an insertion or deletion is called, depending on if the outlier value is lower (deletion) or higher (insertion) than the median. Median is used here rather than mean to provide robustness against extreme outliers. For example, in the case of an extreme outlier due to a deletion from a thriving member in the community, the mean would be skewed and thus could call all three nodes an outlier; whereas the median would take the value of one of the non-deletion nodes and thus, given the two non-deleted nodes carry a similar value, only the deletion would be an outlier. A similar process is conducted to search for complex indels. Here, each square (cycle of length 4) in the graph is searched for outliers. If the square either has a single outlier or two outliers that do not have an edge between them (opposites in the square) and one is greater than the median while the other is smaller, a complex indel is called. A tandem duplicate can be called under two different scenarios. The first, a self-duplicate, shown by an edge log fold change of any self-loop edge greater than 1 for any subsequent steps in the series. The second is the situation where the duplicate produces a second node containing a nearly duplicate sequence and loops between two nodes. This is detected by searching all edges with weight $w \geq 2$ for a log fold change edge weight greater than 1. If these criteria are met, the node with the greater log fold change coverage between the two is then called a tandem duplication if it has not been called for another SV at the specified step.

Experiments

Simulated HGT events

Rhea was compared to the metagenome HGT detection tool MetaCHIP by simulating long reads from the simulated HGT events completed in the HgtSIM manuscript (35). For this community, 10 strains within class *Alphaproteobacteria* and 10 strains within class *Betaproteobacteria* were selected. 1 gene was selected from each *Alphaproteobacteria*, mutated with rate m, and inserted randomly into each *Betaproteobacteria*. This resulted in a total of 100 HGT events for the community (Fig 2a). Three long read metagenomic datasets of 500,000 reads were simulated from these reference genomes with NanoSim (40) v3.1.0 with default parameters: a pre-transfer community (*T0*) of the 20 reference genomes in equal abundance, and two separate post-transfer communities with mutation rate $m = 0$ and $m = 30$ ($T1_{m0}$, $T1_{m30}$), which include the 10 original *Alphaproteobacteria* and the 10 HGT-inserted *Betaproteobacteria* references in varying abundances (Fig 2b). These varying abundances were established by randomly selecting relative quantity between 1 and 5 for each of the species as input into the NanoSim abundance text file. MetaCHIP v.1.10.12 was run with GTDB-Tk (8) v2.2.6 with taxonomy release 207 and `-r` set to class (c). Rhea v1.0 was run with default parameters, metaFlye v2.9.3, and minigraph v0.20. Simulated HGT insertions were mapped against reported HGT sequences for both methods using minimap2 (21) v2.24 with default parameters; each HGT insertion sequence was marked as detected if the sequence had a hit to a reported HGT insertion.

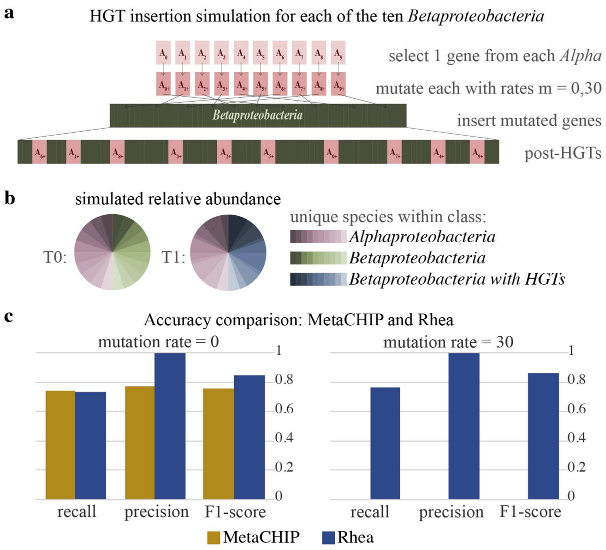
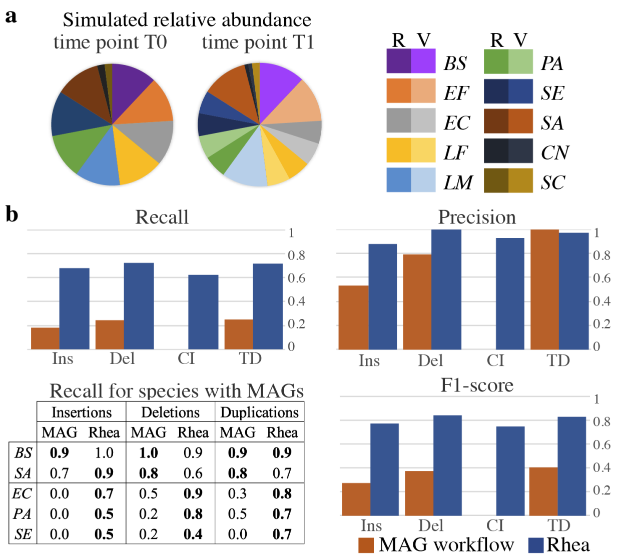


Fig. 2. (a) HGT simulation process completed in the HgtSIM publication (35). One gene is randomly selected from each of the 10 *Alphaproteobacteria* species, mutated with rate m, then inserted into each *Betaproteobacteria*. Mutations rates $m = 0$ and $m = 30$ are included in this study. (b) Simulated relative abundances for time points *T0* and *T1*. *T0* is a simulation of the 20 reference genomes in equal abundance; *T1* is simulated from the 10 original *Alphaproteobacteria* species and the 10 mutated *Betaproteobacteria* species in varying abundances (c) Precision, recall, and F1-score for MetaCHIP (36) and rhea detected insertions for the mock community with mutation rates 0 and 30. Time point *T1* is used for MetaCHIP results; change from *T0* to *T1* is used for rhea.

Simulated SVs

To evaluate the accuracy of rhea for detection of SV types insertion, deletion, complex indel, and tandem duplication in comparison with a MAG-based workflow, variants of each of the 10 microbes in the ZymoBIOMICS Microbial Community Standard were generated. SURVIVOR (15) v1.0.7 was used to randomly create 20 indels (insertions or deletions) and 10 tandem duplicates of length 500-2000 base pairs, with homozygous_ratio=0.5 and Number_haploid=1 in the parameters file, for each of the 10 reference genomes independently. Then a custom script introduced 10 random complex indels of the same length range into each of the variant strains. The custom script randomly selected a location along the genome, then performed a deletion and a random insertion, each within the prescribed length range. Two long read metagenomic datasets of roughly 500,000 reads were simulated from these reference genomes with NanoSim: a pre-transfer community (*T0*) of the original references in their provided relative abundances and a post-transfer community (*T1*), which includes only the variant strain for half of the species and equal abundance of variant and original strains for the other half (Fig 3a). For our MAG workflow, reads were assembled with metaFlye (19) with `--keep-haplotypes` set to true, contigs were binned with MetaBat (17) v2.15 with default parameters, and bins were classified with GTDB-Tk. Bins with the same classification in both simulated samples were analyzed for SVs with MUM & Co (29) v3.8 with the known reference genome length for parameter `-g`. Simulated SV sequences were mapped against reported SV sequences for both methods using minimap2. Each simulated SV was marked as detected if the sequence had a hit to a reported SV sequence with the correct SV type. Since MUM & Co does not call complex indels, we considered these correct if both the deletion sequence and the insertion sequence were returned.



Recall for species with MAGs

	Insertions		Deletions		Duplications	
	MAG	Rhea	MAG	Rhea	MAG	Rhea
BS	**0.9**	1.0	**1.0**	0.9	**0.9**	**0.9**
SA	0.7	**0.9**	**0.8**	0.6	**0.8**	0.7
EC	0.0	**0.7**	0.5	**0.9**	0.3	**0.8**
PA	0.0	**0.5**	0.2	**0.8**	0.5	**0.7**
SE	0.0	**0.5**	0.2	**0.4**	0.0	**0.7**

Fig. 3. (a) Relative abundance of long reads for two simulated time points (*T0*, *T1*) for our ZymoBIOMICS community. Each of the 10 microbes were randomly given 20 indels, 10 tandem duplications, and 10 long complex indels to create a variant strain (15). *T0* contains only the original references (R); *T1* introduces the variants (V), where half the species have variants in equal abundance to their original reference *[Escherichia coli (EC), Lactobacillus fermentum (LF), Pseudomonas aeruginosa (PA), Salmonella enterica (SE), Cryptococcus neoformans (CN)]*, and half the species are dominated by their variants *[Bacillus subtilis (BS), Enterococcus faecalis (EF), Listeria monocytogenes (LM), Staphylococcus aureus (SA), Saccharomyces cerevisiae (SC)]*. (b) Complete recall, precision, and F1-score for each of the SV types (Ins: insertion, Del: deletion, CI: complex indel, TD: tandem duplication) for both workflows (bar plots) and recall on a subset of 5 species (table). For the MAG workflow, MAGs were curated for *T0* and *T1* separately. Then, Mum & Co called SVs between *T0* and *T1* MAGs of matching taxonomic classification. The 5 species selected for the table are the 5 species with a classified MAG at both time points. The top portion (BS, SA) show the species where the variant dominates in *T1*; whereas both the variant and the original reference are present in *T1* for the bottom portion (EC, PA, SE). The better recall is in **bold** for each comparison.

Cheese rind ripening

To evaluate rhea on a real microbiome, PacBio HiFi metagenomic reads from cheese rinds throughout ripening were taken from a previous study (34). One rhea run for “Cheese C” was completed with the 5 corresponding samples in temporal order and parameter `--type` set to pacbio-hifi. The assembly graph connected component that showed interesting evolutionary patterns was classified with GTDB-Tk (8) “classify-wf” with default parameters, and is referred to as the *Halomonas* subgraph per this taxonomic classification. Mobile genetic element (MGE) contigs and putative hosts were established in the original publication utilizing Hi-C sequencing technology, overlap read coverage, and the viralAssociatePipeline (6). To determine which of these contigs showed signatures in our *Halomonas* subgraph, BLAST (2) was run for all MGE contigs with a putative host, against the extracted *Halomonas* subgraph sequences as reference with default parameters. MGE contigs were considered to have their signatures present in the graph if a hit with query coverage $> 5\%$ was reported. One subsection of the *Halomonas* subgraph was selected for further investigate as it showed a change in dominating graph path over time. Nodes within this path were characterized with SeqScreen-Nano (3) v4.1 with default parameters and provided SeqScreen databases v21.4.

Hot spring microbial mat sequencing

Microbial mat plugs were extracted from Mushroom Spring, Yellowstone National Park, USA on July 30, 2009 across a series of temperatures: 50°C, 55°C, 60°C, 65°C. DNA was quantified using the Qubit 3.0 Fluorometric Quantitation dsDNA High Sensitivity kit (ThermoFisher Scientific, Waltham, MA, USA) and stored for future use at -80°C. DNA extractions were analyzed using the Genomic DNA ScreenTape Analysis kit on the 4150 TapeStation System (both from Agilent, Santa Clara, CA, USA). Size selection using AMPure XP beads (Beckman Coulter, San Jose, CA, USA) increased DNA fragment length from a mean of 2kb up to 6kb with high recovery of DNA. Size selected DNA was prepped for sequencing using the Oxford Nanopore Technologies (ONT) 1D Genomic DNA by Ligation library preparation kit (SQK-LSK109, Oxford Nanopore Technologies, Oxford, UK). Libraries were then sequenced using the ONT MinION sequencer using one FLO-MIN106D R9 Version Rev D flow cell per temperature sample. Sequencing was run on a MacBook Pro (model A1502, Apple) using ONT's MinKNOW software. Automatic basecalling through this software was turned off. Sequencing runs lasted between 24-44 hours. Basecalling was completed using the ONT software Guppy (https://github.com/nanoporetech/pyguppyclient.git) with default parameters.

Hot spring microbial mat analysis

Rhea was run on Oxford Nanopore Technologies (ONT) reads from a hot spring microbial mat for 4 unique temperatures (see above) to asses an environmental microbiome with a high-level of complex microbial interactions (5; 26). Basecalled sequences were listed in order of increasing temperature with the `--collapse` parameter set to true. MAGs were also curated for reads from the $60°C$ sample by metaFlye assembly with `--keep-haplotypes` set to true and contigs binned with MetaBat 2 (17). Each read was then aligned back to the set of MAGs with minimap2 with default parameters. Reads with an alignment to a MAG contig of $> 80\%$ of length were considered to be included in MAGs, mimicking the pipeline of a previous manuscript (4). Kraken 2 (39) v2.1.1 was additionally run with the Kraken 2 default parameters and RefSeq indexes released on May 17, 2021 for all raw reads in this sample.

Results

Simulated HGT insertions

Two simulation experiments were conducted with a community of strains within *Alphaproteobacteria* and *Betaproteobacteria* classes to evaluate HGT detection accuracy: one with mutation rates $m = 0$ and the other with $m = 30$. For the HGT insertions with $m = 0$, rhea delivered comparable recall to MetaCHIP (0.73 to 0.74) and improved precision (1.0 to 0.77) (Fig 2c). The only non-insertion SV that rhea called was a single complex indel, which was due to two insertions sequences in close genomic proximity. Given the two inserted sequences were still detected as sequences of increasing abundance, this was still considered this an accurate call. Although results for MetaCHIP and rhea for $m = 0$ were relatively similar, a large discrepancy was observed for mutation rate $m = 30$. Here, the accuracy for rhea stays consistent to that of no mutations (0.76 recall and 1.0 precision), yet MetaCHIP is not able to detect any of the HGT insertions. This caveat is also highlighted in the MetaCHIP manuscript; the inserted sequence is required to be present in another MAG (putative donor) in the community for MetaCHIP to be able to detect the HGT insertion. Additionally, MetaCHIP returned a total of 13 false positive insertions, while rhea did not report any false positives.

Simulated structural variants

A single simulated experiment was conducted to evaluate rhea in comparison to a MAG-based workflow for a variety of SVs. This experiment contained two mock time points (*T0* and *T1*), where *T0* contains only the references in the ZymoBIOMICS Microbial Community Standard and *T1* contains a mix of original references and simulated variants. For the 400 simulated SVs, rhea greatly outperformed the MAG workflow in terms of recall (Fig 3). While rhea detected 71, 68, 63, and 72 of the simulated insertions, deletions, complex indels, and tandem duplications respectively, the MAG workflow only identified 19, 23, 0, and 25, respectively. This discrepancy was largely due to the inability to curate independent MAGs for low abundant species and SV distinctions.

MAGs were classified for 5 of the 10 species at both *T0* and *T1*, limiting the MAG-based workflow to only attempt to call SVs for these species. Of the 5 species, 2 (*B. subtilis, S. aureus*) were from species where the SV-containing strain dominated in sample *T1*, while 3 (*E. coli, P. aeruginosa, S. enterica*) contained both the original and the SV-containing strains in *T1*. Accuracy results between rhea and MAG pipelines proved comparable for insertions, deletions, and tandem duplicates when only the SV-strain was present in post-transfer sample *T1*. However, when both the original and SV-strains were present, only one MAG was curated for the species, leaving many of the SV graph nodes unbinned and thus impossible to detect. Since the SV caller used in the MAG workflow does not call complex indels, we considered a complex indel to be detected if both the insertion and deletion for the complex indel was reported; however, this was not the case for any of the 50. For the two low abundant fungi present in only 2% relative abundance, MAGs were not created at either time point, while rhea was able to detect SVs for these species with similar recall to the more abundant bacteria. Even with the reduced potential to call SVs with a MAG-based workflow, this process resulted in 21 false positive SVs call while rhea only elicit 17.

Of the 125 SVs that were not detected by rhea, roughly 50% were not detected in the assembly graph, roughly 40% were in the graph but resolved into longer nodes rather than partaking in SV graph patterns, and the remaining 10% were called as the wrong SV type.

Cheese ripening temporal series

To demonstrate rhea's ability to extract interesting microbial evolutionary patterns within a microbiome over time, PacBio HiFi metagenomic sequences taken from a cheese rind over the course of ripening were used as input (34). A total of 5 samples were included from sampling weeks 2, 3, 4, 9, and 13, creating 4 pairs of change (*C1-4*). Evaluating the assembly graph coverage visuals produced by rhea and Bandage (38), one connected component stood out for displaying significant graph complexity and diversity in coverage, implying a disproportionately large number of SVs. Rhea SV results indicated roughly 20% of SVs in the community to be contained in this subgraph (Fig 4a). This connected component was then classified by GTDB-Tk under genus *Halomonas* and further exploration was pursued.

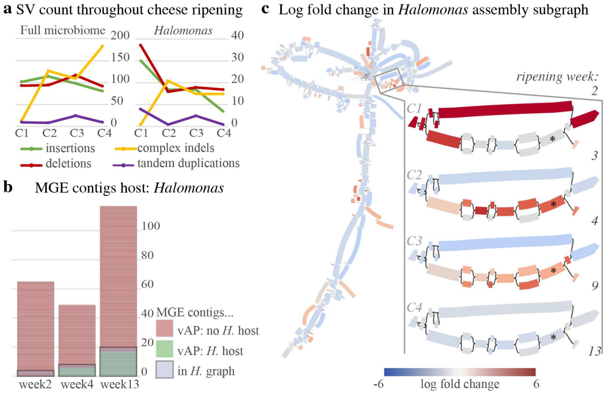


Fig. 4. (a) SV counts detected by rhea for pairs of subsequent samples throughout cheese ripening (*C1-4*) for the entire community and exclusively the extracted *Halomonas* subgraph. (b) Previously established MGE contigs for 3 selected time points, described as either with (green) or without (red) *Halomonas* host by viralAssociationPipeline (vAP) per original publication's findings. Grey boxes signify the MGE contigs that had a BLAST hit of > 5% query coverage to our *Halomonas* subgraph. (c) Rhea and Bandage generated visual for the log fold change in coverage for the *Halomonas* subgraph. Left shows the complete *Halomonas* subgraph between weeks 4 and 9 (*C3*), selected for showing a general decrease in abundance yet an increase in abundance for several subsequences. Right zooms in on a small portion of the subgraph containing an interesting evolutionary pattern, where the log fold change in coverage graph is shown for each pair of subsequent time points (*C1-4*). The graph node marked with a ∗ indicates the node containing the predicted type I restriction-modification system.

First, the ability for viral and plasmid mobile genetic elements (MGEs) to show signatures in the *Halomonas* subgraph was evaluated. In the original publication for the cheese samples, MGE contigs and putative hosts were established via Hi-C sequencing technology and overlap read coverage with the viralAssociatePipeline (6) for sampling weeks 2, 4, and 13. Their results showed *Halomonas* to be host for 0, 6, and 17 MGE contigs, respectively. A BLAST (2) comparison of all MGE contigs against the *Halomonas* subgraph, showed all putative *Halomonas* MGE contigs to display signatures in our *Halomonas* subgraph (hit with more than 5% query coverage), despite previous host connections being defined via Hi-C sequencing and our graph being constructed solely on long-read sequences. An additional 4, 2, and 3 MGE contigs showed signature in the *Halomonas* subgraph without having a previous description of a *Halomonas* host for the time point for each of the 3 included sampling weeks respectively (Fig 4b), which may be false positives or novel host discovery. Finally, one striking section of the *Halomonas* subgraph was selected for gene function analysis (Fig 4c). Here, a newly emerged path (displayed lower option) shows an increase in coverage over time up until stabilizing by week 9, suggesting an evolutionary advantage over the alternative path (top option). Gene function predictions returned by SeqScreen (3) showed the newly dominating path to contain a type I restriction-modification system that was not expressed in the alternative sequence. This suggests an evolutionary advantage due to phage protection in the *Halomonas* strains, which is unsurprising given the increasing number of phage interactions detected throughout ripening for *Halomonas*. Exploratory analysis here demonstrates a novel approach produced by rhea to extract genomic subsequences that suggest an evolutionary advantage, gain insight into MGE hosts, and infer microbial interactions.

Hot spring microbial mat temperature series

Lastly, to assess an environmental sample with complex interactions, rhea was run on a temperature series of samples taken from the Mushroom Spring microbial mat in Yellowstone National Park, USA. Samples were collected from 4 different portions of the mat with temperatures 50°C, 55°C, 60°C, and 65°C. Rhea detected SVs between subsequent temperature increases (Table 2). An extraordinarily large number of SVs were detected in the hot spring microbial mat, averaging 8.9 million per consecutive pair, as opposed to an average of 317 per pair in the cheese microbiome. The vast quantity of SVs is particularly noticeable for complex indels, as counts for this type was observed to be over an order of magnitude greater than the other SV types observed. The number of detected complex indels increased with the first two temperature increases (over 8 million and 22 million, respectively), but then fewer are detected with the last temperature increase (over 3 million). While this decrease implies more stability at these higher temperatures, a closer look at the coassembly graph and alignments could confirm this pattern is true signal rather than a result of decreased average read length in the 65°C sample. Previous research closely analysed two *Synechococcus* isolates from these mats and showed a large number of diverse insertion sequence (IS) activity occurring within the two strains (26). Our findings suggest there is far more transposon and gene exchange occurring in microbial mats that has yet to discovered, and likely many uncharacterized novel bacterial strains. Further research is needed to confirm these suspicions and additionally detect the gene functions for the thriving SVs to give insight into evolutionary drivers for these extremophiles.

Table 2. Sample and SV statistics for hot spring microbial mat temperature series. SV counts shown represent the number of SV detected between the sample listed in the row and the previous row. SV types abbreviations are as follows: Ins: insertion, Del: deletion, TD: tandem duplication, and CI: complex indel.

Sample	Reads (million)	Bps (billion)	Ins (10^3)	Del (10^3)	TD (10^3)	CI (10^3)
50°C	3.6	7.5				
55°C	2.4	6.2	224	232	0.21	8616
60°C	3.4	7.7	220	239	0.38	22217
65°C	2.9	3.7	212	242	0.19	3611

One sample (60°C) was selected to assess read inclusion rate of alternative workflows for this community rife in unknown microbes. To evaluate a reference-based taxonomic classification method, reads were classified by Kraken2 with default database, where 42% of the reads were left unclassified. To evaluate a MAG creation workflow, MAGs were created with MetaFlye contigs and MetaBat2 binning, where roughly 30% of raw reads did not map to a binned contig. Use of rhea allowed for the inclusion of all sequenced reads to distinguish subsequences and genomic context specific to high temperature environments and give insight into the evolutionary history of these active and uncharacterized microbes within hot spring microbal mats.

Computational usage

All software analysis was completed on a Ubuntu 22.04 LTS system with 15 threads. The `/usr/bin/time` command was used to gather time and memory statistics. Reported CPU (central processing unit) time was calculated by summing the user and the system time; RAM (random access memory) requirements were determined using the maximum resident set size.

Table 3. Computational usages for rhea experiments.

study	reads (million)	base pairs (billion)	User+sys time (h)	RAM (GB)
HgtSIM (m0)	1.0	4.0	13	26
HgtSIM (m30)	1.0	4.0	13	26
ZymoBIOMICS	1.0	4.0	13	26
Cheese	1.8	23.1	154	47

Discussion

Here we present rhea, a novel method for detecting structural variants (SVs) between consecutive samples in long-read metagenome series data. Rhea leverages sequence information from the entire metagenomic community and avoids need for a reference database or MAG creation by analyzing structural motifs and change in alignment coverage on a combined coassembly graph. This permits SV detection for intra-species variations, low abundance genomes, and novel organisms. Our simulated results of recent HGT events and SVs in two mock communities show rhea to outperform existing methods. Use of rhea on a cheese rind microbiome with samples taken throughout ripening allowed us to infer MGE hosts that align with Hi-C sequencing and additionally suggest recently transferred genes with a suspected evolutionary advantage for the host. Use of rhea for a varying temperature series of samples from a hot spring microbial mat allowed us to include reads that would likely have been removed in alternative workflows, as strain-level diversity prevents sequences from being incorporated in MAGs and lack of isolate reference genomes prevent use of reference-based approaches. While extracting evolutionary insights from this complex community still provides a significant challenge, rhea introduces a first step in logically parsing these metagenomic sequences.

Methods for identifying significant changes throughout a metagenome series is an active area of research (43). Currently, a common approach is to first simplify each metagenome into a profile that can be logically aligned and compared, such as taxonomic classification relative abundance, gene function presence, and counts of short sub-sequences (k-mers) (10). Yet, each of these strategies either oversimplifies potentially important sequences of microbial communities or is biased by a reference database (20; 25). Rhea results contain input data for the interactive visual software package Bandage (38), for exploration of changes in graph coverage throughout a metagenome series. This tool provides researchers with an efficient method to investigate sequence-level fluctuations while maintaining genome context, to ultimately extract sequences of interest (Fig 4c). It is important to note that metagenomic sequences simply provide a snapshot of the microbial community at the time of sampling, and thus oscillating fluctuations that take place between samples may not be detected.

Currently, rhea is only able to detect insertion, deletion, tandem duplication, and complex indel SV types between two metagenomes of similar microbes. The method could theoretically be expanded to inversions and translocations, however, we anticipate the need to maintain node directionality (whether the sequence is read forward or reverse) in the evaluated coassembly graph. Rhea could also be expanded to detect more complex patterns of multiple overlapping SVs or short read sequences, but further experimentation is required.

Rhea has so far only been evaluated for SV detection over the course of microbiome series data. The idea of constructing a coassembly graph and comparing the coverage between samples could be expanded beyond series data and used for different types of studies, such as cohort comparison analyses and MGE host detection. As the number of reads included in the study increases, methods of downsampling sequences to generate the graph or an alternate graph construction methods could be considered. Alternative graphs, could also be explored in attempt to improve sensitivity for SV detection, given that results in our mock ZymoBIOMICS community still collapsed nearly a quarter of simulated SVs. However, alternative graph structures could also create two unique connected components for microbes that have undergone significant structural variations, which would prevent the current detection algorithm within rhea to call such SVs. Further analysis could help determine at which diversity levels SVs are collapsed in a single node or separated into unique connected components, to provide genome similarity requirement guidelines for SV detection capability within rhea.

In lieu of metagenome-specific methods, metagenomes are often construed to fit methods and models developed for genome analyses. Yet this simplification overlooks inherent complexities of dynamic and interdependent microbial ecosystems (7). By viewing these communities holistically and acknowledging their intricate interplay and co-evolution, we can discover nuanced patterns, novel relationships, and a deeper understanding of the collective behaviors throughout the community. Developed to embody this ideology, rhea is a novel technique to pinpoint microbial heterogeneity and evolution by capturing the full essence of these diverse and interconnected ecosystems.

Data availability

Rhea and all associate code are available on GitHub (https://github.com/treangenlab/rhea). Scripts, simulations, complete results, and hot spring long reads are available on OSF under project FVHW8.

Competing interests

No competing interest is declared.

Author contributions statement

K.D.C. derived the concept, developed the software, completed experiments and wrote the manuscript. S.E.V. conducted sequencing. B.Y., D.B., S.S., R.C., E.P.C.C. and T.J.T. contributed to method development. All authors read, revised and approved the manuscript.

Acknowledgments

K.D.C. was supported in part by Ken Kennedy Institute Computational Science and Engineering Graduate Recruiting Fellowship, Rice University Wagoner Foreign Study Scholarship, and the Chateaubriand Fellowship. T.J.T. was supported in part by NIH grant P01-AI152999 from the National Institute of Allergy and Infectious Diseases (NIAID) and NSF CAREER award IIS-2239114 (PI Treangen). K.D.C., S.S., and T.J.T. were supported by the NSF MIM Universal Rules of Live (URoL) grant (EF-2126387, PI Treangen). This project has received funding from the European Union's Horizon 2020 research and innovation programme under Marie Skłodowska-Curie grant agreement Nos. 872539 and 956229 (R.C.). R.C. was supported by ANR Transipedia, SeqDigger, Inception and PRAIRIE grants (ANR-18-CE45-0020, ANR-19-CE45-0008,PIA/ANR16-CONV-0005, ANR-19-P3IA-0001). D.B. acknowledges support from a BBSRC-NSF/BIO collaborative research grant (award number 1921429), NSF_Emerging Frontiers grant (award number 2125965); Joint Genome Institute Community Sequencing Project # 503441 and # 509352, and the Carnegie Institution for Science. The work conducted by the U.S. Department of Energy Joint Genome Institute, a DOE Office of Science User Facility, is supported by the Office of Science of the U.S. Department of Energy. S.E.V. is supported by NSF GRFP (fellow ID 2023333162).

References

1. Abante, J., Wang, P.L., Salzman, J.: DIVE: A reference-free statistical approach to diversity-generating and mobile genetic element discovery. Genome Biology **24**(1), 240 (Oct 2023). https://doi.org/10.1186/s13059-023-03038-0
2. Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J.: Basic local alignment search tool. Journal of Molecular Biology **215**(3), 403–410 (Oct 1990). https://doi.org/10.1016/S0022-2836(05)80360-2
3. Balaji, A., Liu, Y., Nute, M.G., Hu, B., D. Kappell, A., S. Lesassier, D., D. Godbold, G., Ternus, K., Treangen, T.: SeqScreen-Nano: A computational platform for streaming, in-field characterization of microbial pathogens. In: Proceedings of the 14th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics. pp. 1–10. BCB '23, Association for Computing Machinery, New York, NY, USA (Oct 2023). https://doi.org/10.1145/3584371.3612960
4. Benoit, G., Raguideau, S., James, R., Phillippy, A.M., Chikhi, R., Quince, C.: High-quality metagenome assembly from long accurate reads with metaMDBG. Nature Biotechnology pp. 1–6 (Jan 2024). https://doi.org/10.1038/s41587-023-01983-6
5. Bhaya, D., Grossman, A.R., Steunou, A.S., Khuri, N., Cohan, F.M., Hamamura, N., Melendrez, M.C., Bateson, M.M., Ward, D.M., Heidelberg, J.F.: Population level functional diversity in a microbial community revealed by comparative genomic and metagenomic analyses. The ISME journal **1**(8), 703–713 (Dec 2007). https://doi.org/10.1038/ismej.2007.46
6. Bickhart, D.M., Watson, M., Koren, S., Panke-Buisse, K., Cersosimo, L.M., Press, M.O., Van Tassell, C.P., Van Kessel, J.A.S., Haley, B.J., Kim, S.W., Heiner, C., Suen, G., Bakshy, K., Liachko, I., Sullivan, S.T., Myer, P.R., Ghurye, J., Pop, M., Weimer, P.J., Phillippy, A.M., Smith, T.P.L.: Assignment of virus and antimicrobial resistance genes to microbial hosts in a complex microbial community by combined long-read assembly and proximity ligation. Genome Biology **20**(1), 153 (Aug 2019). https://doi.org/10.1186/s13059-019-1760-x
7. Brito, I.L.: Examining horizontal gene transfer in microbial communities. Nature Reviews Microbiology **19**(7), 442–453 (Jul 2021). https://doi.org/10.1038/s41579-021-00534-7
8. Chaumeil, P.A., Mussig, A.J., Hugenholtz, P., Parks, D.H.: GTDB-Tk v2: Memory friendly classification with the genome taxonomy database. Bioinformatics **38**(23), 5315–5316 (Dec 2022). https://doi.org/10.1093/bioinformatics/btac672
9. Chen, L., Zhao, N., Cao, J., Liu, X., Xu, J., Ma, Y., Yu, Y., Zhang, X., Zhang, W., Guan, X., Yu, X., Liu, Z., Fan, Y., Wang, Y., Liang, F., Wang, D., Zhao, L., Song, M., Wang, J.: Short- and long-read metagenomics expand individualized structural variations in gut microbiomes. Nature Communications **13**(1), 3175 (Jun 2022). https://doi.org/10.1038/s41467-022-30857-9
10. Curry, K.D., Nute, M.G., Treangen, T.J.: It takes guts to learn: Machine learning techniques for disease detection from the gut microbiome. Emerging Topics in Life Sciences p. ETLS20210213 (Nov 2021). https://doi.org/10.1042/ETLS20210213
11. Durrant, M.G., Bhatt, A.S.: Microbiome genome structure drives function. Nature microbiology **4**(6), 912–913 (Jun 2019). https://doi.org/10.1038/s41564-019-0473-y
12. Fonstein, M., Haselkorn, R.: Physical mapping of bacterial genomes. Journal of Bacteriology **177**(12), 3361–3369 (Jun 1995). https://doi.org/10.1128/jb.177.12.3361-3369.1995
13. Ghurye, J., Treangen, T., Fedarko, M., Hervey, W.J., Pop, M.: MetaCarvel: Linking assembly graph motifs to biological variants. Genome Biology **20**(1), 174 (Aug 2019). https://doi.org/10.1186/s13059-019-1791-3
14. Ghurye, J.S., Cepeda-Espinoza, V., Pop, M.: Metagenomic Assembly: Overview, Challenges and Applications. The Yale Journal of Biology and Medicine **89**(3), 353–362 (Sep 2016)
15. Jeffares, D.C., Jolly, C., Hoti, M., Speed, D., Shaw, L., Rallis, C., Balloux, F., Dessimoz, C., Bähler, J., Sedlazeck, F.J.: Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. Nature Communications **8**(1), 14061 (Jan 2017). https://doi.org/10.1038/ncomms14061
16. Jiang, X., Hall, A.B., Arthur, T.D., Plichta, D.R., Covington, C.T., Poyet, M., Crothers, J., Moses, P.L., Tolonen, A.C., Vlamakis, H., Alm, E.J., Xavier, R.J.: Invertible promoters mediate bacterial phase variation, antibiotic resistance, and host adaptation in the gut. Science (New York, N.Y.) **363**(6423), 181–187 (Jan 2019). https://doi.org/10.1126/science.aau5238

17. Kang, D.D., Li, F., Kirton, E., Thomas, A., Egan, R., An, H., Wang, Z.: MetaBAT 2: An adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. PeerJ **7**, e7359 (Jul 2019). https://doi.org/10.7717/peerj.7359
18. Kerkvliet, J.J., Bossers, A., Kers, J.G., Meneses, R., Willems, R., Schürch, A.C.: Metagenomic assembly is the main bottleneck in the identification of mobile genetic elements. PeerJ **12**, e16695 (Jan 2024). https://doi.org/10.7717/peerj.16695
19. Kolmogorov, M., Bickhart, D.M., Behsaz, B., Gurevich, A., Rayko, M., Shin, S.B., Kuhn, K., Yuan, J., Polevikov, E., Smith, T.P.L., Pevzner, P.A.: metaFlye: Scalable long-read metagenome assembly using repeat graphs. Nature Methods **17**(11), 1103–1110 (Nov 2020). https://doi.org/10.1038/s41592-020-00971-x
20. LaPierre, N., Ju, C.J.T., Zhou, G., Wang, W.: MetaPheno: A critical evaluation of deep learning and machine learning in metagenome-based disease prediction. Methods (San Diego, Calif.) **166**, 74–82 (Aug 2019). https://doi.org/10.1016/j.ymeth.2019.03.003
21. Li, H.: Minimap and miniasm: Fast mapping and de novo assembly for noisy long sequences. Bioinformatics **32**(14), 2103–2110 (Jul 2016). https://doi.org/10.1093/bioinformatics/btw152
22. Li, H., Feng, X., Chu, C.: The design and construction of reference pangenome graphs with minigraph. Genome Biology **21**(1), 265 (Oct 2020). https://doi.org/10.1186/s13059-020-02168-z
23. Li, Y., Cao, J., Wang, J.: MetaSVs: A pipeline combining long and short reads for analysis and visualization of structural variants in metagenomes. iMeta **2**(4), e139 (2023). https://doi.org/10.1002/imt2.139
24. Liu, R., Zou, Y., Wang, W.Q., Chen, J.H., Zhang, L., Feng, J., Yin, J.Y., Mao, X.Y., Li, Q., Luo, Z.Y., Zhang, W., Wang, D.M.: Gut microbial structural variation associates with immune checkpoint inhibitor response. Nature Communications **14**(1), 7421 (Nov 2023). https://doi.org/10.1038/s41467-023-42997-7
25. Nayfach, S., Pollard, K.S.: Toward Accurate and Quantitative Comparative Metagenomics. Cell **166**(5), 1103–1116 (Aug 2016). https://doi.org/10.1016/j.cell.2016.08.007
26. Nelson, W.C., Wollerman, L., Bhaya, D., Heidelberg, J.F.: Analysis of Insertion Sequences in Thermophilic Cyanobacteria: Exploring the Mechanisms of Establishing, Maintaining, and Withstanding High Insertion Sequence Abundance ∇. Applied and Environmental Microbiology **77**(15), 5458–5466 (Aug 2011). https://doi.org/10.1128/AEM.05090-11
27. Nijkamp, J.F., Pop, M., Reinders, M.J.T., de Ridder, D.: Exploring variation-aware contig graphs for (comparative) metagenomics using MaryGold. Bioinformatics **29**(22), 2826–2834 (Nov 2013). https://doi.org/10.1093/bioinformatics/btt502
28. Ochman, H., Lawrence, J.G., Groisman, E.A.: Lateral gene transfer and the nature of bacterial innovation. Nature **405**(6784), 299–304 (May 2000). https://doi.org/10.1038/35012500
29. O'Donnell, S., Fischer, G.: MUM&Co: Accurate detection of all SV types through whole-genome alignment. Bioinformatics **36**(10), 3242–3243 (May 2020). https://doi.org/10.1093/bioinformatics/btaa115
30. Quince, C., Nurk, S., Raguideau, S., James, R., Soyer, O.S., Summers, J.K., Limasset, A., Eren, A.M., Chikhi, R., Darling, A.E.: STRONG: Metagenomics strain resolution on assembly graphs. Genome Biology **22**(1), 214 (Jul 2021). https://doi.org/10.1186/s13059-021-02419-7
31. Rocha, E.P.C.: Neutral Theory, Microbial Practice: Challenges in Bacterial Population Genetics. Molecular Biology and Evolution **35**(6), 1338–1347 (Jun 2018). https://doi.org/10.1093/molbev/msy078
32. Rocha, E.P.: Order and disorder in bacterial genomes. Current Opinion in Microbiology **7**(5), 519–527 (Oct 2004). https://doi.org/10.1016/j.mib.2004.08.006
33. Roerink, S.F., van Schendel, R., Tijsterman, M.: Polymerase theta-mediated end joining of replication-associated DNA breaks in C. elegans. Genome Research **24**(6), 954–962 (Jun 2014). https://doi.org/10.1101/gr.170431.113
34. Saak, C.C., Pierce, E.C., Dinh, C.B., Portik, D., Hall, R., Ashby, M., Dutton, R.J.: Longitudinal, Multi-Platform Metagenomics Yields a High-Quality Genomic Catalog and Guides an In Vitro Model for Cheese Communities. mSystems **8**(1), e00701–22 (Jan 2023). https://doi.org/10.1128/msystems.00701-22
35. Song, W., Steensen, K., Thomas, T.: HgtSIM: A simulator for horizontal gene transfer (HGT) in microbial communities. PeerJ **5**, e4015 (Nov 2017). https://doi.org/10.7717/peerj.4015
36. Song, W., Wemheuer, B., Zhang, S., Steensen, K., Thomas, T.: MetaCHIP: Community-level horizontal gene transfer identification through the combination of best-match and phylogenetic approaches. Microbiome **7**(1), 36 (Mar 2019). https://doi.org/10.1186/s40168-019-0649-y
37. West, P.T., Chanin, R.B., Bhatt, A.S.: From genome structure to function: Insights into structural variation in microbiology. Current Opinion in Microbiology **69**, 102192 (Oct 2022). https://doi.org/10.1016/j.mib.2022.102192
38. Wick, R.R., Schultz, M.B., Zobel, J., Holt, K.E.: Bandage: Interactive visualization of de novo genome assemblies. Bioinformatics **31**(20), 3350–3352 (Oct 2015). https://doi.org/10.1093/bioinformatics/btv383
39. Wood, D.E., Lu, J., Langmead, B.: Improved metagenomic analysis with Kraken 2. Genome Biology **20**(1), 257 (Nov 2019). https://doi.org/10.1186/s13059-019-1891-0
40. Yang, C., Chu, J., Warren, R.L., Birol, I.: NanoSim: Nanopore sequence read simulator based on statistical characterization. GigaScience **6**(4), gix010 (Apr 2017). https://doi.org/10.1093/gigascience/gix010
41. Ye, K., Wang, J., Jayasinghe, R., Lameijer, E.W., McMichael, J.F., Ning, J., McLellan, M.D., Xie, M., Cao, S., Yellapantula, V., Huang, K.l., Scott, A., Foltz, S., Niu, B., Johnson, K.J., Moed, M., Slagboom, P.E., Chen, F., Wendl, M.C., Ding, L.: Systematic Discovery of Complex Indels in Human Cancers. Nature medicine **22**(1), 97–104 (Jan 2016). https://doi.org/10.1038/nm.4002
42. Zeevi, D., Korem, T., Godneva, A., Bar, N., Kurilshikov, A., Lotan-Pompan, M., Weinberger, A., Fu, J., Wijmenga, C., Zhernakova, A., Segal, E.: Structural variation in the gut microbiome associates with host health. Nature **568**(7750), 43–48 (Apr 2019). https://doi.org/10.1038/s41586-019-1065-y
43. Zhou, B., Wang, C., Putzel, G., Hu, J., Liu, M., Wu, F., Chen, Y., Pironti, A., Li, H.: An integrated strain-level analytic pipeline utilizing longitudinal metagenomic data (Feb 2022). https://doi.org/10.1101/2022.02.15.480548