De novo assembly of microbial genomes from human gut metagenomes using barcoded short read sequences

Eli L. Moss[1,*] & Alex Bishara[2,*], Ekaterina Tkachenko[1], Joyce B. Kang[1], Tessa M. Andermann[3], Christina Wood[4], Christine Handy[4], Hanlee Ji[4], Serafim Batzoglou[2,**], Ami S. Bhatt[1,**]

[1] Departments of Genetics and Medicine, Stanford University, Stanford, California, USA.

[2] Department of Computer Science, Stanford University, Stanford, California, USA.

[3] Department of Medicine, Division of Infectious Diseases, Stanford University, Stanford, California, USA.

[4] Department of Medicine, Division of Oncology, Stanford University, Stanford, California, USA.

[*] These authors contributed equally to the study.

[**] To whom correspondence should be addressed: asbhatt@stanford.edu or serafim@cs.stanford.edu

Abstract

Shotgun short-read sequencing methods facilitate study of the genomic content and strain-level architecture of complex microbial communities. However, existing methodologies do not capture structural differences between closely related co-occurring strains such as those arising from horizontal gene transfer and insertion sequence mobilization. Recent techniques that partition large DNA molecules, then barcode short fragments derived from them, produce short-read sequences containing long-range information. Here, we present a novel application of these short-read barcoding techniques to metagenomic samples, as well as Athena, an assembler that uses these barcodes to produce improved metagenomic assemblies. We apply our approach to longitudinal samples from the gut microbiome of a patient with a hematological malignancy. This patient underwent an intensive regimen of multiple antibiotics, chemotherapeutics and immunosuppressants, resulting in profound disruption of the microbial gut community and eventual domination by *Bacteroides caccae*. We significantly improve draft completeness over conventional techniques, uncover strains of *B. caccae* differing in the positions of transposon integration, and find the abundance of individual strains to fluctuate widely over the course of treatment. In addition, we perform RNA sequencing to investigate relative transcription of genes in *B. caccae*, and find overexpression of antibiotic resistance genes in our *de novo* assembled draft genome of *B. caccae* coinciding with both antibiotic administration and the appearance of proximal transposons harboring a putative bacterial promoter region. Our approach produces overall improvements in contiguity of metagenomic assembly and enables assembly of whole classes of genomic elements inaccessible to existing short-read approaches.

Introduction

Short-read metagenomic sequencing and assembly have played an instrumental role in advancing the study of bacterial genomes beyond the minority of culturable organisms[1]. This has greatly expanded our understanding of the genomic structure and dynamics of the human microbiome, which has emerged as an important element in many aspects of health and disease[2–4]. However, the precise genetic makeup of organisms within these complex systems remains poorly understood.

Duplicated and conserved sequences within a metagenome complicate the recovery of strain-level architecture with existing short-read methods. These sequences arise from several mechanisms, including horizontal gene transfer and transposon mobilization, each with a well-described capacity to induce significant changes in phenotype. Horizontal gene transfer (HGT) results in the acquisition and dissemination of functional elements that can include antibiotic resistance genes, virulence factors, or metabolic capabilities[5,6]. Mobile elements can affect gene function and regulation by disrupting coding sequences[7], or by introducing new promoter sequences[8,9], and have been observed to mobilize in response to antibiotic stress[10]. In addition, gene upregulation mediated by mobile sequences has been shown to increase antibiotic resistance[8,11]. HGT and mobile sequence element duplication represent two mechanisms by which bacterial genomes acquire repeated sequences, which pose challenges for metagenomic sequence analysis.

Specialized computational techniques have been developed to recover draft genomes for individual organisms within metagenomic samples. These techniques include dedicated metagenomic assemblers[12–14] and contig binning approaches based on sequence similarity[15–18] and coverage depth covariance[19]. Binning techniques can group assembled contigs into significantly more comprehensive drafts, but do not improve the contiguity of the input assembly. Sequence fragment sizes from existing high throughput platforms are too short to span duplicated sequences, and as a result, these regions currently remain unassembled. This represents an inherent limitation of short reads, necessitating a complementary molecular approach to assemble these classes of genomic sequences.

In principle, long-read approaches can be used to address these issues. Single molecule platforms such as Pacific Biosciences' Single Molecule Real Time sequence approach have been successfully applied to close genomes of cultured isolates[20–22] and dominant organisms within more complex mixtures[23], as well as assemble mobile and duplicated sequences within cultured isolates[24,25]. Synthetic long reads have also been used to improve metagenomic

assemblies[26]. However, limited throughput and relatively high input DNA mass requirements bar these approaches from application to biological samples where high molecular weight DNA is limited. In addition, single molecule approaches have comparatively low nucleotide accuracy, which may impede resolution of closely related strains.

Several existing platforms address these shortcomings by partitioning long input DNA fragments, and barcoding shorter fragments derived from them, to tag short reads with long-range information. The recent 10X Genomics platform streamlines this barcoding process with more than 100,000 droplet partitions to yield uniquely barcoded short-read fragments from one or a few long molecules trapped in each droplet partition[27]. Sequencing of 10X libraries yields shallow coverage depth groups of barcode-sharing reads, which we will refer to as read clouds[28] (also referred to as linked reads[27]). This platform offers an attractive combination of high nucleotide accuracy, low input mass requirements and long-range information. Applications of this platform and similar ones predating it have focused on reference-based human haplotype phasing[26,27,29–31], and their potential for *de novo* metagenomic sequence assembly has yet to be explored.

We present, to our knowledge, the first approach that leverages read clouds provided from the 10X Genomics Gemcode platform to directly assemble complex metagenomic mixtures from a single sample. We developed an assembler, Athena, that uses the barcode information in order to assemble sequences that cannot be placed in correct genomic context using short reads alone. We first tested our approach on a mock mixture of DNA from 10 known bacterial species and used Athena to accurately assemble and place multiple copies of the ribosomal RNA (rRNA) operon within the draft assembly of each species.

We then applied our technique to a clinical gut microbiome time series from a patient undergoing stem cell transplantation for a hematological malignancy. This patient underwent an extensive series of antibiotic, antiviral, antifungal, chemotherapeutic and immunosuppressive treatments, imparting profound selective pressure on the gut microbiome, resulting in domination by *Bacteroides caccae*. Our barcoded assembly approach reveals the presence of a number of nearly identical *B. caccae* strains differing in the position of integration of transposons and a large transferred region (genomic island), which we validated using long-range PCR and Sanger sequencing. Our improved drafts allowed us to quantify gene expression in regions that were fragmented or unresolved by short-read assembly, revealing potential transcriptional effects of insertion sequences and larger mobile elements not otherwise accessible by conventional short-read assembly approaches.

Results

Read cloud sequencing and Athena Assembly

We developed the Athena assembler to use long-range information encoded within barcoded short-read sequences. In our approach, we first extract high molecular weight DNA and use the 10X Genomics Gemcode platform to obtain barcoded short reads for our samples (Figure 1a). Barcoded short reads are first jointly assembled with a conventional short-read assembler (see Methods) to obtain an initial covering of the metagenome in the form of assembled sequence contigs. These seed contigs are then provided to the Athena assembler for further metagenome sequence assembly (Figure 1b). The same barcoded short reads are then mapped back to the seed contigs and read pairs that span contigs are used to form edges in a scaffold graph. Branches in this scaffold graph correspond to ambiguities encountered by the short-read assembler. At each edge, Athena examines the short-read mappings together with the attached barcodes to propose a simpler subassembly problem of a pooled subset of barcoded reads that can potentially assemble through branches in the scaffold graph (see Supplementary Methods). The selection of this read subset removes the majority of reads considered during the initial assembly while retaining reads that cover the local target sequence, isolating the local subassembly problem from the broader metagenome. The much smaller and independent subassembly problems are performed for every edge in the scaffold graph to yield longer, overlapping subassembled contigs that resolve branches in the scaffold graph. The initial seed contigs and intermediary subassembled contigs are then passed to an overlap layout consensus assembler, Canu[32] (formerly the Celera Assembler), which determines how to assemble the target genome from these much longer contigs. The resulting metagenome assembly consists of more complete sequence contigs with resolved repeats that are too difficult to assemble with short-read techniques alone.

Assembly of highly conserved elements in a synthetic metagenomic community

As a first validation of our approach, we performed read cloud sequencing on a mock mixture of DNA from 10 known bacterial species (see Methods) and tested Athena's ability to accurately assemble the conserved 16S and 23S ribosomal RNA operon subunits. This repeat

unit, which varies in size from 5kb to 7kb depending on the size of the spacer between 16S and 23S subunits, is known to occur in multiple nearly identical copies throughout individual bacterial genomes[33,34]. These subunits are also highly conserved across all bacterial species, yet contain sufficient divergence between species, allowing for their use as an informative marker for phylogenetic characterization of microbial communities[35]. These interspecies repeat units, which also occur in multiplicity within each individual genome, serve as a useful model of the assembly issues created by duplicated and conserved sequences.

Our validation process entailed Athena assembly, 16S/23S operon identification, and molecular validation of assembled ribosomal RNA loci. We produced both standard short-read and read cloud libraries for our ten species mixture, and applied conventional assembly and Athena on each of these libraries, respectively. To obtain drafts for each of the ten organisms for each approach, we classified resultant contigs, and grouped contigs sharing species-level classifications (see Methods). We used RNAmmer[36] to find instances of rRNA subunits within the short-read and Athena drafts. Conventional short-read assembly produced a single disconnected instance of the rRNA operon most closely resembling that of *Bacteroides ovatus*. In contrast, Athena read cloud assembly assembled 28 copies of the complete rRNA operon (Supplementary Table 1). In addition, Athena assembled 2 copies of 16S and 2 copies of 23S outside the operon context. Of the 32 total instances of assembled rRNA, we chose 23 with at least 3kb of assembled sequence on left and right flanks for long-range PCR validation, allowing flanking PCR primer design for amplification across the entire repeat (Supplementary Figure 1). All 23 PCR experiments produced specific amplicons of the anticipated length. To further validate rRNA assemblies, we obtained Sanger reads confirming the left and right junctions between the operon and genomic flanking sequence (Supplementary File 1) for those instances of the rRNA operon occurring within the Klebsiella genome. Sanger primer sequences targeted regions of high nucleotide identity internal to the operon, and initiated Sanger read extension outward into genomic flanks (Supplementary Figures 2,3). In summary, we demonstrated validated assembly of multiple copies of highly conserved rRNA subunits in a mock mixture of bacterial DNA, and were able to show markedly improved capability over conventional short-read assembly to resolve the genomic contexts of these sequences.

Assembly of a clinical gut microbiome time series

To test the generalizability of this approach to natural biological samples, we next applied Athena to longitudinal clinical gut microbiome samples obtained from a patient receiving treatment for hematological malignancy at the Stanford Hospital Blood and Marrow Transplant Unit. The patient underwent hematopoietic stem cell transplantation (HCT) for myelodysplastic syndrome and myelofibrosis, which was refractory to treatment with azacitidine. The patient received multiple medications during the period of observation, including antibiotics, antivirals, antifungals, chemotherapeutics and immunosuppressives. The patient was diagnosed with gastrointestinal (GI) graft-versus-host disease (GVHD) (clinical grade 3, histological grade 1 in duodenal biopsy). Fiber, fermented foods, probiotics, and prebiotics were restricted from the patient's diet over the entire time series. During the course of treatment, the patient's gut microbiome underwent profound simplification, rapidly becoming dominated by *Bacteroides caccae*, a rare opportunistic pathogen[37] with mucin-degradation capability[38] (Figure 2; for genus-level classifications see Supplementary Figure 4; for classification data see Supplementary File 4).

To study the trajectory of this patient's gut microbiome throughout this treatment, we selected the following four time points for sequencing: **A** (day 0) pre-chemotherapy and pre-HCT; **B** (day 13) post-chemotherapy and post-HCT; **C** (day 43) post broad-spectrum antibiotic exposure and onset of GI GVHD; **D** (day 56) long-term follow up (Figure 2). We prepared both Illumina Truseq and 10X Gemcode libraries for stool samples from these four time points, equalized read counts between the two library types, and assembled each library with a short-read assembler and Athena, respectively (see Methods). We classified the resulting contigs from each metagenomic assembly, and grouped those contigs sharing species-level classifications to obtain drafts of each constituent organism for each approach.

We observed significant improvements in both contiguity and completeness for several of the drafts produced by the read cloud approach as compared to standard short-read techniques (Supplementary Table 2). Our approach yielded drafts with improved contiguity in 15 of 16 organisms with >30x coverage depth in both libraries (average 4.8-fold increase in N50). These belonged to the genera *Bacteroides, Parabacteroides,* and *Haemophilus* . These Athena drafts were as complete as those obtained by short-read assembly (average completeness percentages within 2%), ascertained by core gene detection (see Methods, Supplementary File 2). However, improved contiguity in Athena drafts allowed an additional 7.6Mbp of assembled sequence to be assigned to these 16 organisms.

We found that per-species read coverage differed substantially between standard and read cloud libraries, with five organisms receiving >30x coverage in only one of the two libraries (Supplementary Table 2). These discrepancies in coverage are likely due in part to biases introduced by either fragment size selection or sampling effects during the microfluidic molecular partitioning process. We observed an approximate 10-fold decrease in DNA mass during size selection prior to read cloud library preparation, potentially causing depletion of more fragmented organism genomes. In order to separate effects on assembly performance arising from upstream molecular steps from those intrinsic to the Athena assembly approach itself, we examined the 16 organisms that were sufficiently covered in both libraries (>30x short read coverage).

For the predominant *Bacteroides caccae*, Athena was able to consistently yield more contiguous (time points A, B, C, D) and more complete (time points B, C, D) drafts than conventional techniques (Figure 3). Our best draft in time point D had an N50 of 390kb and total size of 5.5Mb, as compared to an N50 of 61kb and total size of 4.5Mb yielded by short-read techniques. Given these much-improved drafts, we next sought to locate duplicated sequences resolved by Athena. We posited that comparative analysis of the *B. caccae* drafts across time points may yield insight into either selection or potential genomic remodeling that this organism may have undergone as it grew to eventually dominate this host's gut microbiome.

Read clouds recover nearly identical strains in clinical samples

In order to locate duplicated sequences that were assembled by Athena, we identified short *k*-mers that were overrepresented in the Athena assemblies relative to the short-read assemblies, and annotated the most highly overrepresented sets with BLAST (nr/nt database)[39]. Of the elements thus identified, we focus on mobile element IS612, a conserved Bacteroides insertion sequence (IS), due to its high prevalence in our time series and pronounced fluctuation in abundance across the time series. This sequence is present in the short-read assemblies, but only appears in a single copy with extreme sequence coverage depth (up to 17,345x coverage in time point C) that is detached from genomic context, highlighting a major limitation of standard short-read assembly.

From our read cloud assemblies we selected 44 distinct instances of the IS from all timepoints. Selected assemblies had at least 3kb assembled on both flanks, which served as primer design sites for validation by long-range PCR and Sanger sequencing. We were able to

obtain Sanger sequencing data validating the specific genomic placement of all but one of the 44 instances. Of the 43 validated IS instances, 20 occurred in contigs classified as *B. caccae* (Figure 3). The other validated instances were also within contigs classified as belonging to *Bacteroides*: six in *B. vulgatus,* two each in *B. thetaiotaomicron, B. ovatus, and B. dorei*, and a single instance each in *B. uniformis* and *B. xylanisolvens*. The remaining 10 did not receive species-level classifications.

At many insertion sites in *B. caccae*, short-read alignments to the Athena assembly confirmed the co-occurrence of both strains harboring the IS and strains with the pre-insertion ancestral sequence. From these alignments, we obtained an estimate of the relative abundance of ancestral and insertion-containing strains for each site (see Methods). These estimated abundances agreed with observed PCR band intensities at the ancestral and IS-containing band sizes (Figure 4). We observed large shifts in the ancestral abundance at several insertion loci (Supplementary Table 3), with 18 large (>50%) shifts occurring between consecutive time points.

In addition to these small-scale structural divergences, we observed larger-scale structural strain variations with similarly pronounced shifts in abundance in *B. caccae*. Inspection of the global alignments revealed one instance of the IS to be immediately adjacent to a large 60kb region present in only some strains in the last time point. Raw short-read alignments from C and D to the Athena draft from D confirmed this rearrangement and showed the 60kb sequence to be present at much lower abundance relative to flanking genomic sequence during time point C (Figure 5). Annotation of this 60kb island revealed the presence of xerC and xerD tyrosine recombinases, which can mediate genomic integration of mobile elements[40,41]. In addition, the region contains flagellar motor protein motB, a phage integrase family protein, and an operon encoding four genes mediating streptomycin biosynthesis. We searched for xerCD recognition motifs previously described in *Escherichia coli*[40] within our draft from time point D and found a single 11 base pair site directly adjacent to the island and overlapping with an IS. PCR validation confirmed the integration of the 60kb region, as well as the pre-integration strain containing only the adjacent IS (Supplementary Figure 5). We have demonstrated the validated assembly of numerous instances of a small mobile element as well as a large-scale sequence integration with extensive functional potential, and observed these assembled sequences to fluctuate widely in relative abundance over time.

Identification of insertion-mediated transcriptional upregulation

To explore Athena's effects on metatranscriptomic data analysis in genomic regions where short-read assemblies would be otherwise too fragmented, we used our drafts as references in RNA sequence data alignment. We performed RNA sequencing on time points B, C, and D of the same samples (RNA yield from sample A was very poor, despite multiple attempts). We compared use of both short-read and Athena drafts as references, and found our more complete *B. caccae* drafts allowed a significantly larger fraction of all RNA sequencing reads to be assigned to this organism. Specifically, an additional 11%, 22%, and 10% of the RNA sequencing reads from time points B, C, and D respectively were aligned to the Athena drafts of *B. caccae* over the corresponding short-read drafts. Our more complete drafts allowed for a much larger fraction of the coding potential of this organism to be evaluated.

We next used our Athena drafts together with the RNA sequencing reads to investigate the potential transcriptional effects of the structural changes we detected, focusing on IS612. This IS contains a putative outward-facing promoter near its 5' end oriented antisense to its transposase coding sequence[8]. Determining the transcriptional effect of this IS is difficult in a complex metagenomic setting, as RNA-seq reads may originate from co-occurring strains with or without a given insertion. In light of this difficulty, we restricted our attention to integration sites that were dominated first by ancestral strains and then by IS-harboring strains in consecutive time points, with at least 30% change in estimated ancestral abundance. In these sites, a corresponding increase in transcription downstream of the promoter versus upstream transcription is more likely attributable to the additional promoter provided by the IS. We located three such genomic loci of "transcriptional asymmetry", all demonstrating more than 10-fold higher transcription of the downstream neighboring gene relative to upstream (Figure 6, Supplementary Figures 6 and 7, Supplementary Table 4).

The highest degree of transcriptional asymmetry coincided with placement of the putative promoter in IS612 to upregulate norM, a multidrug resistance transporter (Figure 6a). NorM is a multidrug efflux protein found to confer resistance to ciprofloxacin[42], which was administered for the first 30 days of treatment through time points A and B (Figure 2). Short-read alignments to this insertion site showed this integration to be undetectable in time point A, present in roughly a third of strains in B, and then in the majority in time points C and D, consistent with visible band patterns in our targeted PCR results (Figure 6b, Supplementary Table 3). Other transcriptional asymmetries were observed in susC, an outer membrane protein involved in starch utilization[43], and resA, an oxidoreductase involved in cytochrome c synthesis[44]

(Supplementary Figures 6, 7). Short-read alignments showed these insertions to be absent in this site in time point A, and present in roughly a third of the strains in B, the majority of strains in time point C, and half of strains in time point D. The abrupt changes observed in the abundance of these insertions suggest strong selective pressures.

We found the most highly expressed gene in time points C and D to be the extended-spectrum beta-lactamase gene per1, known to confer resistance to meropenem, a major component of the patient's antibiotic regimen. This gene was expressed nearly 60% more than the second most expressed gene in both time points C and D (Supplementary Table 4). Its rise in expression coincided with a two-week course of meropenem beginning 10 days after time point B and ending 6 days before time point C. Per1 continues to exhibit high expression in time point D, 19 days after withdrawal of meropenem, despite the absence of any further beta-lactam antibiotic administration. Our drafts located an IS adjacent to this gene oriented correctly for IS-mediated transcription to occur. While estimating relative abundance of this insertion, we were unable to detect reads from the ancestral strain, and determined this insertion to be fixed within the population over the course of treatment.

Though several insertion sequences became undetectable in DNA sequence data between timepoints C and D, strains with insertions adjacent to norM, susC and per1 continue to dominate through the end of our time course. By time point D, ciprofloxacin and meropenem have been withdrawn for 26 and 19 days, respectively, yet expression of resistance factors norM and per1 remained increased compared to levels prior to antibiotic exposure.

Discussion

We present a novel approach for applying barcoded short reads to metagenomics. This method improves contiguity in assembled drafts and enables assembly of whole classes of duplicated elements inaccessible to conventional short-read approaches. The improved draft assemblies generated by Athena serve as a useful basis for RNA sequence data alignment that allow us to investigate transcriptional effects of newly assembled sequences. In clinical samples, we show that the powerful combination of genome assembly with gene expression analysis yields evidence for antibiotic resistance mechanisms that may evolve quickly via insertion sequence mobilization. We observe apparent transcriptional upregulation of antibiotic resistance and starch metabolism genes by adjacent mobile sequence elements, highlighting the clinical importance of regulatory changes induced by these duplicated sequences.

After confirming the accuracy and evaluating the performance characteristics of this approach using a defined bacterial mock community, we proceeded to apply this approach to human microbiome samples. The clinical subject we examine is an individual who underwent extensive treatment with several classes of medication in a short time period while undergoing hematopoietic stem cell transplantation. Immune suppression and extensive antibiotic treatment rendered the patient studied in this investigation highly susceptible to destabilization and taxonomic simplification of the gut microbiome, which has been associated with increased overall mortality, GI GVHD and other complications[4,45–47]. Past work shows that antibiotic use and a restricted diet can lead to intestinal domination by one or few microorganisms[38,48], but the mechanisms by which any specific microorganism achieves dominance over the larger community remain poorly understood.

Barcode assembly of our clinical gut microbiome time series reveals that populations of microbes with apparently stable taxonomic composition can be composed of many closely related strains undergoing wide fluctuations in abundance. Though our later two time points C and D have nearly identical species compositions, they each harbor numerous strains differing in mobile sequence integration sites. In addition, the most dominant strains in time point D carry a 60kb island containing 72 predicted coding sequences including a four-gene operon encoding streptomycin biosynthesis. The capability of strains to acquire novel genetic material or alter regulation of existing genes generates strain diversity, potentially increasing the likelihood that some highly fit strain evolves to dominate the destabilized gut microbiome of immunocompromised hosts.

IS mobilization in particular provides a means by which organisms can upregulate individual or sets of genes past levels of endogenous expression. *In vitro* experiments have confirmed that insertion sequences similar to IS612 are mobilized in response to antibiotic stress[10] and during strain competition[49]. These elements can upregulate adjacent genes, potentially resulting in increased antibiotic resistance[8]. In our clinical time series, insertions expected to affect genes related to antibiotic resistance and starch utilization reached the highest abundances within a timespan of days, consistent with antibiotic use and potentially related to dietary interventions. Our results confirm the occurrence of this conserved mechanism in the gut microbiome, and suggest that it functions as an adaptive response mechanism underlying clinically significant alterations in bacterial behavior.

Our approach to *de novo* assemble sample-specific strain diversity complements existing approaches to analyze strain-level variation. Previous methods use compiled reference sequence collections, and characterize nucleotide divergence within defined gene sets or gene presence within an organism-specific pan-genome[50–52]. Although these methods differ substantially, their shared reliance on the reference sequence collection restricts sensitivity in the context of poorly sequenced or unknown species, comprising a large fraction of microbial diversity[53], or previously unobserved structural variants of known species. Our results provide a means to obtain significantly improved individual drafts from metagenomes that could enable existing tools to study strain-level variation that is currently underrepresented in existing reference sequence collections.

We anticipate that the approach presented here, which enables short-read assembly to capture new classes of duplicated sequences, will benefit from future improvements in molecular barcoding platforms. In our clinical metagenomes, Athena was unable to produce improved drafts for some organisms due to disproportionately reduced read representation in read cloud library preparation compared to standard methods. We anticipate that improvements in high molecular weight DNA extraction, which preserve relative abundance between species, will allow Athena to provide improved drafts for these organisms as well. We also anticipate that improvements in both coverage uniformity and long fragment partitioning during read cloud library preparation will enable our approach to eventually produce near reference grade microbial genomes from individual gut microbiome samples alone.

Methods

Mock community DNA mixture preparation

For the mock community, purified bacterial isolate DNA was obtained from BEI Resources (Manassas, VA) for each of ten species belonging to distinct genera (Supplementary Table 1). DNA samples were diluted to 5ng/uL and combined in equal volumes. The mixture was concentrated by ethanol precipitation and resuspended in 1/4x volume, and quantified with the Thermo Qubit 3.0 fluorometer (Thermo Fisher Scientific, Waltham, MA) using the Qubit dsDNA HS assay. The mixture was size selected with the Sage Science BluePippin instrument (Sage Science, Beverly, MA) with the BLF7510 agarose gel cassette kit and Marker S1 targeting a 5kb minimum fragment length, then quantified again with Qubit prior to library preparation.

Clinical participant recruitment

The patient from whom the longitudinal stool samples were obtained was recruited at the Stanford Hospital Blood and Marrow Transplant Unit under an IRB-approved protocol (PIs: Dr. David Miklos, Dr. Ami Bhatt). Informed consent was obtained. A comprehensive chart review was carried out to identify clinical features of the patient, demographic information, duration and exposure to medications, and diet.

Clinical DNA and RNA preparation

Stool samples were obtained from the study subject on an approximately weekly basis, when available. Clinical stool samples were placed at 4C immediately upon collection, and processed for storage at -80C the same day. Stool samples were aliquoted into 2mL cryovial tubes with either no preservative or 700uL of RNAlater and homogenized by brief vortexing. Samples were stored at -80C until extraction. DNA was extracted from stool samples with the Qiagen QiAMP Stool Mini Kit modified with the addition of 7 cycles of 30s bead beating alternating with 30s cooling on ice (for full details, see Supplementary Methods). DNA concentration estimations were performed using Qubit fluorometric quantitation. Extracted DNA was size selected with the BluePippin instrument with a 5kb minimum size cutoff as described

above. DNA for the synthetic mixture and clinical samples was prepared for sequencing with the Gemcode instrument (10X Genomics, Pleasanton, CA) with revision C of the standard protocol. Library fragment size was quantified with the 2100 Bioanalyzer instrument (Agilent Technologies, Santa Clara, CA) using the High Sensitivity DNA chip and reagent kit.

RNA was extracted with the RNeasy Mini kit (Qiagen, Germantown, MD) from samples stored in RNAlater at -80C. Total RNA concentration was assayed with the Qubit RNA HS kit and Qubit fluorometer. RNA was ethanol precipitated and resuspended in nuclease-free water to concentrate, then quantified again using both Qubit RNA HS and Qubit DNA HS kits to determine the degree of DNA contamination. Contaminating DNA was removed using the Baseline-ZERO DNase protocol (Epicentre, Madison, WI) with 30 minute incubation followed by a second ethanol precipitation. Ribosomal RNA was depleted with the Epicentre Ribo-Zero rRNA removal kit (Bacteria). The rRNA-depleted RNA was quantified with 2100 Bioanalyzer using the Agilent RNA6000 Pico kit. cDNA sequencing library was prepared with the Illumina (San Diego, CA) Truseq Stranded mRNA kit following the Truseq Stranded mRNA LS protocol.

In addition, conventional short-read sequencing libraries were prepared for DNA extracted from clinical stool samples, the mock mixture and unmixed bacterial isolates. Clinical samples were prepared for sequencing with the Illumina Truseq Nano DNA library preparation kit with a target insert size of 500bp. Mixed and unmixed bacterial isolate DNA samples were prepared for sequencing with the Illumina Nextera XT library preparation kit.

Sequencing and assembly

Bacterial isolate libraries and 10x Gemcode libraries were subjected to 2x148bp sequencing with the Illumina Nextseq 500. The mock mixture, clinical DNA and clinical RNA libraries were subjected to 2x100bp sequencing with the Illumina HiSeq 4000. Raw reads from conventional sequencing were trimmed using Trim Galore v0.4.1[54] using a minimum length of 60bp, minimum terminal base score of 20, and the Illumina adapter sequences. In addition, forward reads were trimmed by 2bp at the 5' end and reverse reads were trimmed by 6bp at the 3' end to remove low quality bases, and deduplicated with SuperDeduper v2.0[55]. All short-read libraries were downsampled to equal the total read count of the corresponding read cloud library. Data were assembled using SPAdes v3.6.1[56] with default parameters for paired-end input. SPAdes seed assemblies of Gemcode libraries were then reassembled with Athena. Assemblies were visualized with IGV[57], R[58] and python using the ggplot2[59], circlize[60] and

matplotlib[61] libraries. *K*-mers enriched in Athena assemblies compared to conventional assemblies were determined using Jellyfish[62] for *k*-mer counting.

Assembly annotation

Contigs were assigned taxonomic classifications using Kraken[63] with a custom database constructed from the Refseq and Genbank[64,65] bacterial genome collections. All species represented in <1% of read classifications were discarded. Sequences were functionally annotated using Prokka v1.11[66].

Insertion abundance estimation

Illumina Truseq short read data were aligned with BWA[67] to the validated insertion sequence assemblies obtained from Athena assemblies of clinical microbiome data. Reads recruited to each insertion locus were realigned with STAR[68] in order to obtain gapped alignments spanning the insertion sequence. Gapped alignments, representing the ancestral strain, were counted for each insertion. To obtain relative abundance, ancestral counts were divided by coverage sampled at a location two kilobases adjacent to the left flank of the insertion.

PCR amplification

PCR was performed to establish molecular contiguity between sequences assembled by Athena, and to generate template materials for Sanger sequencing. PCR reactions contained Phusion High-Fidelity DNA Polymerase (New England BioLabs, Ipswich, MA) with Phusion HF Buffer and NEB Deoxynucleotide Solution Mix. Primers were obtained from Elim Biopharm (Hayward, CA) with target melting temperature of 60˚C. Protocols for rRNA and IS amplifications were as follows:

Per 30uL reaction:

- 6uL 5x HF buffer
- 0.6uL 10mM dNTP
- 0.3uL Phusion
- 2uM 10mM forward primer

- 2uM 10mM reverse primer
- 1uL template DNA
- PCR clean water to 30uL

Thermocycling:

1. Denature: 98˚C for 30s
2. 35 cycles
    a. Denature: 98˚C for 5s
    b. Anneal: 65˚C for 10s
    c. Extend: 72˚C for 30s/kb
3. Final extension: 72˚C for 5min
4. Hold: 4˚C indefinitely

Sanger sequence assembly validation

In order to validate the Athena assembly at locations likely to be misassembled, we targeted duplicated sequences for orthogonal molecular validation. In the mock mixture, we identified occurrences of the 16S/23S operon within the Klebsiella genome, and in the clinical samples, we isolated all assembled regions containing insertion sequence IS612. For each of these regions, we designed PCR primers adjacent to the left and right junctions between the duplicated sequence and genomic flank. We amplified each genomic segment, performed gel extraction to isolate the amplicon corresponding to the inserted variant of the genomic segment when necessary, and performed Sanger sequencing traversing the left and right junctions. For very low abundance insertions, nested PCR using primers surrounding the left and right junctions of the insertion assembly was performed in order to amplify sufficient material for sequencing. Primer sequences and Sanger read sequences can be found in Supplementary File 1. Primer design and Sanger sequencing data visualization, quality control and alignment were performed with Geneious v7.1.4[69].

Code availability

The Athena assembler together with a demonstration dataset can be found at https://github.com/abishara/athena_meta. This example closes several gaps within an initial

draft assembly from SPAdes, including assembling two instances of IS612 inside a single contig of *B. caccae*.

Data availability

The datasets generated during the current study are available in the NCBI Sequence Read Archive under Bioproject accession PRJNA380276.

Acknowledgements

The authors would like to thank Alexandra Sockell for assistance operating the NextSeq 500, and Arend Sidow for valuable feedback on the manuscript. This work was supported by NCI K08 CA184420, the Amy Strelzer Manasevit Award from the National Marrow Donor Program, and a Damon Runyon Clinical Investigator Award to A.S.B. E.L.M. was supported by National Science Foundation Graduate Research Fellowship DGE-114747. A.B. was supported by the Stanford Genome Training Program (SGTP; NIH/NHGRI) and the Training Grant of the Joint Initiative for Metrology in Biology (JIMB; NIST). Access to shared compute resources was supported in part by NIH P30 CA124435 using the Stanford Cancer Institute Shared Resource Genetics Bioinformatics Service Center.

Author contributions

E.L.M., A.B., A.S.B. and S.B. conceived of the study. E.L.M., C.H., C.W. and H.J. prepared read cloud libraries. E.L.M., E.T., J.K., and T.A. collected samples, extracted DNA and RNA and prepared short-read sequencing libraries. E.L.M. designed and selected samples, and performed read cloud sequencing, PCR validation, and Sanger sequencing. A.B. and S.B. conceived of the assembly approach. A.B. implemented the Athena assembler. E.L.M. and A.B. carried out all analyses, wrote the manuscript, and generated figures. All authors commented on the manuscript.

Competing financial interests

The authors declare no competing financial interests.

Figure Legends

Figure 1

Overview of read cloud library prep and Athena Assembly for metagenomes.

- **a)** DNA is first extracted from clinical stool samples and size selected to enrich high molecular weight (HMW) DNA. HMW DNA then undergoes 10X Genomics Gemcode library preparation to sparsely partition small numbers of long DNA fragments across >100,000 droplet partitions. Degenerate amplification of these long fragments is then performed within these partitions to obtain barcoded short-fragment libraries. Each short-fragment has P3 and P5 sequence adapters with a barcode unique to its partition. Short-fragments are then pooled and sequenced with an Illumina instrument.
- **b)** The Athena assembler uses read clouds to yield more complete drafts that accurately place repeated sequences. An example repeat that is resolved and placed by Athena is shown in orange. 1) The resulting 10X reads are first assembled with standard short-read techniques to obtain seed contigs, input reads are mapped back to these seeds, and spanning read pairs are used to build a scaffold graph containing unresolvable branches. 2) At each edge, Athena proposes a much simpler subassembly problem on a pooled subset of barcoded reads informed by the scaffold graph mappings. Example barcodes in red and blue are passed to a short-read assembler to perform subassembly for this branch to yield longer contigs that disambiguate branches in the scaffold graph. 3) The resulting subassembled contigs, together with the initial seeds, are then passed to an overlap consensus layout (OLC) assembler that assembles the target from these much longer overlapping pieces. The resulting draft assembly metagenome produces more complete and more contiguous drafts in which repeats are also resolved.

Figure 2

Patient gut microbiome composition and drug exposure during treatment. The study subject was admitted to Stanford Cancer Center with myelodysplastic syndrome and myelofibrosis and subsequently underwent hematopoietic stem cell transplantation (HCT, denoted by a red line). Stool samples were collected prior to HCT and over the following five weeks as the patient underwent chemotherapy, antibiotic treatment, and profound immunosuppression. Taxonomic classification of conventional (Illumina TruSeq Nano DNA) sequencing reads reveals

pronounced dysbiosis emerging following HCT, with gut domination by *Bacteroides caccae*, a rare opportunistic pathogen with mucin degradation capability. Abundances are given as read fractions of data classified at the species level (for genus-level classifications, see Supplementary Figure 2).

Figure 3

Short-read and read cloud drafts of *Bacteroides caccae.* Athena's draft (blue) has improved contiguity and repeat sequence placement compared to the standard short-read draft (dark grey) in *Bacteroides caccae* across four longitudinal clinical metagenomic samples. Tracks are ordered chronologically (A through D) from the outermost inward, with a portion enlarged for detail. The Athena draft has superior contiguity and coverage as compared to that of short reads (Athena: 390kb N50, 5.5Mbp total; short reads: 61kb N50, 4.5Mbp total in time point D). The Athena draft includes 20 total assembled instances of the duplicated IS (closed and open circles) and all three copies of the rRNA operon (green triangles). Both the IS and the rRNA operon fail to assemble contiguously to both genomic flanks using conventional sequencing and assembly techniques, but are assembled properly with the Athena read cloud approach. The reference for alignment consists of contigs classified as *B. caccae* in the Athena assembly of timepoint C. Reference contigs below the N75 size of 87.9kb (25% of total sequence), contig alignments below 2.5kb, and gaps under 1kb are omitted for visual clarity. Contigs from all assemblies other than the Athena assembly of time point C are lighter in color if they aligned to the reference, but were not classified as belonging specifically to *B. caccae.*

Figure 4

Validation of IS and ancestral *B. caccae* strains.

a) Alignments of short reads from time points B, C, and D to a representative IS integration site reveal domination of the ancestral strain without the IS in time points in B and D, and domination of the strain harboring the insertion in C. Short-read alignments from B and D show many ancestral reads (red, indicating global alignment to the ancestral sequence) spanning over both left and right junctions, while short-read alignments from C show many reads supporting the IS integration (blue, indicating read pairs or single reads spanning into the IS).
b) PCR primers were designed to amplify the flanking sequence surrounding the putative insertion site. PCR amplification of the resulting amplicons across all four time points

display short and long bands corresponding to ancestral and IS strains respectively. Relative band intensities agree with quantitative measurements of ancestral read alignments.

Figure 5

Detection of a horizontally transferred 60kb island adjacent to an IS. A 60kb island was found within the Athena assembled genome of *B. caccae* in time point D, but absent from time point C in the clinical longitudinal series. Coverage of short reads from C and D aligned to the Athena draft from D support the absence of this island in time point C, and its presence in time point D. Evidence of the presence of the IS is also absent from earlier time points A and B. The island contains 72 predicted coding sequences including xerC and xerD tyrosine recombinases, previously described to mediate genomic plasmid integration. In addition, the region contains a four-gene operon encoding enzymes involved in streptomycin synthesis. The fragmented reconstruction by short reads does not allow for accurate classification of the majority of the island sequence at the species level. Known recognition sites of the xerC and xerD recombinases were searched across the entire draft, and only a single 11 base pair instance was found directly next to the right junction of the island incorporation. Seven of the 11 base pairs originate from the IS sequence, strongly suggesting that the integration of this IS created a recognition site by which another integration mechanism was able to further mediate rearrangement.

Figure 6

Metagenomic RNA sequencing supports IS-mediated transcription within *B. caccae.* IS612 contains a putative promoter that may affect transcription of neighboring genes. Comparison of RNA sequencing read depths of genes upstream and downstream from this promoter can give insight into its relative contribution to transcription. In later time points C and D, dominant strains harbor an introduced promoter positioned to upregulate norM. In prior time point B, dominant strains have no IS in this region.

a) In time point B, which is dominated by ancestral strains without the promoter, RNA sequencing read coverage depth is relatively equal on both sides of the integration site. In time point C, which is dominated by strains with the introduced promoter, the coverage of the downstream gene norM is much higher than the upstream yidC. The

coverage depths of all neighboring genes increases in time point C relative to B, but this increase is 10-fold greater in downstream genes. Read pairs spanning between the IS promoter and norM further support the transcriptional contribution of the IS. This difference in coverage, and domination by strains with this promoter, both persist through time point D.

b) PCR with primers flanking the above integration instance of IS612 yields amplicons without the insertion sequence in earlier time points A and B (400bp), and with the insertion in later time points C and D (1.9kb).

Supplementary Figure and Table Legends

Supp. Figure 1

PCR amplification of Athena-assembled instances of the rRNA operon, as well as 16S and 23S sequences occurring outside of an operon, from the 10 species mock bacterial mixture. All instances of rRNA sequences that assembled with at least 3kb of flanking sequence on each side were targeted for amplification. Anticipated amplicon sizes are 3.1-3.3kb for 16S-only assemblies, 4.6-4.8kb for 23S-only assemblies, and 6.4-6.7kb for full operon assemblies. Successful amplification at the correct anticipated size was obtained for all tested assemblies. Ladder is Thermo Fisher 1kb Plus DNA Ladder, gel is 1% agarose in TAE buffer. For primer sequences and rRNA operon assemblies, see Supplementary File 1. Lanes 1-2: *Campylobacter upsaliensis*, 16S-only. Lanes 2-3: *Campylobacter upsaliensis*, 23S-only. All following lanes are full rRNA operon assemblies including both 16S and 23S. Lanes 5-8, 12, 22: *Citrobacter freundii*. 9-10, 14, 17, 20: *Klebsiella sp. 1_1_55*. 11, 18, 21: *Enterococcus faecalis*. 13, 16: *Finegoldia magna.* 15, 19: *Bifidobacterium sp. 12_1_47BFAA*. 23: *Acinetobacter radioresistens*.

Supp. Figure 2

Design of PCR and Sanger sequencing primers used in rRNA assembly validation. A multiple alignment was performed of the five *Klebsiella sp. 1_1_55* rRNA operon assemblies chosen for Sanger validation. Primers for PCR amplification were designed in the flanking regions for whole-operon amplification, with target amplicon sizes of approximately 7kb (primers indicated by green lines below each assembly). Primers for Sanger sequencing were designed in high sequence identity regions internal to the operon (indicated by green triangles above Identity track). The locations of 5S, 23S and 16S ribosomal subunit sequences are indicated in separate tracks.

Supp. Figure 3

Example Sanger sequence validation of a *Klebsiella* rRNA operon. For each assembled rRNA operon chosen for Sanger validation, the entire operon and several hundred bp flanking sequence were amplified by PCR (see Supplemental Figure 6). The amplicon was submitted for Sanger sequencing with outward-facing primers targeting conserved regions internal to the operon (indicated by green lines, and in Supplementary Figure 6). Sanger sequences were aligned to the operon assembly and checked for sequence identity and correct alignment.

Supp. Figure 4

Genus level taxonomic composition of the clinical time series, based on classification of Illumina Truseq short read data. For visual clarity, the top ten most abundant genera are shown. Reads receiving no classification are shown ("Unclassified"). Remaining reads are classified at broader taxonomic levels than genus, and are omitted.

Supp. Figure 5

PCR validation of the 60kb transferred island. Lanes 1-4: PCR amplification of a region spanning the left junction between island and genomic flank in time points A, B, C and D, respectively. Lanes 5-8: PCR amplification spanning from left genomic flank across island to right genomic flank. Band sizes are consistent with detection of ancestral strain lacking island. While amplification may have also begun in strains harboring the island, the amplicon is too long for complete extension. The right junction between island and genomic flank is validated by the Sanger sequencing read covering the left junction of the adjacent IS. Ladder is Thermo Fisher 1kb Plus DNA Ladder.

Supp. Figures 6 and 7

Metagenomic RNA sequencing supports IS-mediated transcription within *B. caccae.* Examples of IS-mediated transcription for two additional genes, susC (Supp. Figure 6) and resA (Supp. Figure 7), which are downstream of a promoter introduced by the IS. Comparison of RNA sequencing read depths of genes upstream and downstream from this promoter can give insight into its relative contribution to transcription. In both instances in the later time points C and D, dominant strains harbor an introduced promoter to upregulate susC and resA. In prior time point B, dominant strains have no IS in this region in both instances. The coverage depths of all neighboring genes increases in time point C relative to B, but this increase is 10-fold greater in downstream genes. The transcriptional contribution of the IS is further supported by the presence of read pairs spanning between the IS promoter and adjacent downstream genes.

Supp. File 1

Sanger sequencing validation of assembled rRNA (mock mixture) and insertion sequence (clinical samples). Template sequences for each (assembled rRNA operon instances.fasta, klebsiella rRNA sequences.fasta, insertion_sequences_with_flanks.fa), PCR

primers used (insertion sequence clinical samples pcr primers.tsv, rRNA mock mix primers.fasta) and Sanger sequence reads (sanger_sequences clinical insertion sequence.fasta, sanger_sequences mock mix 16S klebsiella.fasta) are provided.

Supp. File 2

Complete raw output from CheckM[70] for draft genome assemblies obtained from short read and read cloud sequencing of clinical microbiome samples. Output describes draft genome completeness ascertained from lineage-specific core gene set representation within the draft sequence.

Supp. File 3

Read-level classification of clinical samples. Kraken[63] was used to classify Illumina Truseq short read data from clinical samples sequenced on the Illumina HiSeq 4000 (see methods). Figure 2 and Supplementary Figure 2 are based on these classifications. All taxonomic classifications occurring in at least 1% of reads are presented. Column format is as follows:

1. Percentage of reads classified within this taxon (and all child taxa)
2. Number of reads classified within this taxon (and all child taxa)
3. Number of reads classified specifically to this taxon (occurs when a more specific classification is not possible)
4. Letter code indicating classified taxon (Unclassified, Domain, Kingdom, Phylum, Class, Order, Family, Genus, or Species).
5. NCBI taxonomy identifier
6. Classified taxon

Supp. Table 1

Per-species counts of assembled instances of 16S or 23S outside of an operon, as well as counts of assemblies of the entire rRNA operon, assembled by Athena within the bacterial DNA mixture. Counts of the rRNA operon including both 16S and 23S (spades-rrna and athena-rrna columns), and either 16S or 23S alone (spades-16S/athena-16S or spades-23S/athena-23S columns) are given for instances assembled with at least 10bp and 3kb of flanking sequence on both sides. Grant total refers to total occurrences of 16S, 23S, or operons including both within each library type.

Supp. Table 2

Assembly statistics (N50, N80, total bases) in each time point for all bacterial genomes that were covered by at least 30x in either read cloud or standard libraries. Ralstonia is highlighted in red as a contaminant and five organisms that were not covered by at least 30x in both libraries are highlighted in orange. The remaining 16 organisms have >30x coverage in both read cloud and standard libraries. Columns: athena/SR-n50/80: assembly N50/N80; athena/SR-size: total assembled bases per organism; athena/SR-cov: average coverage of assembly per organism; better-n50: names which assembly demonstrated a higher N50; better-sz: names which assembly demonstrated a higher total assembled length.

Supp. Table 3

Fractional abundances of ancestral strains at each assembled insertion site. Read alignments at each validated insertion site allow quantification of strains with the pre-insertion sequence. Reads originating from strains without a given IS instance are recognized by having an alignment with a single gap spanning over the assembled insertion sequence. Abundance of ancestral strains is expressed as a fraction of overall read coverage, sampled two kilobases adjacent to the insertion. Also noted are adjacent upregulated genes mentioned in the main text.

Supp. Table 4

Coding sequence locations for *B. caccae* together with RNA sequencing read counts. Gene annotations were obtained using Prokka and contigs classified as belonging to *B. caccae* from the Athena assembly in time point C were used as the reference when aligning all RNA sequencing reads from time points B, C, and D. The target gene downstream of the promoter as well as the upstream gene are highlighted (green: downstream, red: upstream) for the three integration sites that display IS-mediated upregulation.

10X Gemcode Contamination

Athena assembly yielded a draft for *Ralstonia pickettii* in addition to the 10 intended species for the read cloud library. This organism was present in all read cloud libraries and absent from all conventional short-read libraries prepared from clinical samples, mixed isolate and unmixed isolate DNA. Thus, we attributed these contigs to DNA contamination introduced during 10X Genomics Gemcode library preparation and discarded them.

Athena Assembly

We developed Athena to use barcoded short-read sequences derived from partitioned long input DNA fragments, which we refer to as read clouds. We apply Athena to a read cloud dataset generated with the 10X Genomics Gemcode instrument. In principle, the long fragments that are used as input to these platforms allow resolution of repeats contained within these fragments. However, the barcode-specific coverage of each long fragment is too sparse to allow *de novo* assembly of each in isolation. Furthermore, the long range information encoded within the raw output of each barcode in the form of unordered and unoriented short-read sequences does not fit well into existing sequence assembly algorithms. Athena uses the barcode information to propose a series of simplified assembly tasks that can be performed using existing assemblers as black box subroutines.

Athena first uses an existing short-read assembler (SPAdes) to obtain an initial sequence covering of the underlying metagenome in the form of (possibly short) sequence contigs. A scaffold graph is then constructed using the paired-end information from short-read alignments to these contigs. This scaffold graph contains branches that can be attributed to nearly identical repeats, small divergent sequences between otherwise identical strains, or conserved sequences. Mappings of the barcoded short reads to this scaffold graph allow the selection of input read subsets for a smaller assembly problem (subassembly), such that the resulting contigs yield unambiguous paths through the scaffold graph. The resulting contigs are then passed as reads to an overlap layout consensus assembler, Canu (formerly Celera), for further assembly of these much larger sequences.

The steps for Athena assembly are as follows:

1) A conventional short-read assembler (SPAdes [56]) is used to assemble the raw reads to obtain an initial covering of the target metagenome in the form of short sequence contigs. We refer to the contigs as seeds.

2) Raw reads are mapped back to these seed contigs, and paired end mappings that span two seed contigs are considered for edge creation in a scaffold graph. However, we observed a significant fraction of read pairs mapping with an intervening distance that exceeds the expected library fragment size. We believe these pairs to be mostly due to chimeric

fragments arising during Gemcode library preparation. In order to prevent these from introducing spurious connections in the scaffold graph, we perform the following steps:

2a) For any two seed contigs that are still connected by at least three spanning read pairs, the mapped positions of these spanning read pairs on each seed contig are clustered together into 500bp neighborhoods, corresponding to the average library fragment size.

2b) All clusters are examined and if any single cluster on each seed contig contains more than 50% of these spanning read pairs, then an edge is added in the scaffold graph. Otherwise, the candidate edge is assumed to be spurious and discarded. This filtering process greatly reduces the number of proposed subassemblies to perform.

3) For each remaining edge between two seed contigs within the scaffold graph, subassembly of the linked seed contigs is performed with the following steps:

3a) Barcodes containing at least one read mapping to both seeds are selected. We refer to these as subassembly barcodes. Pooled reads from the subassembly barcodes potentially contain contiguous sequences that bridge together the two seed contigs.

3b) Pooled reads that map to these two seeds are used to estimate short-read coverage of the target sequence within the subassembly. If the short-read coverage is estimated to be low (<10x), then this subassembly is skipped as the local target is unlikely to assemble at low depths. If the short-read coverage is estimated to be high (>200x), then the subassembly barcodes are first downsampled to accelerate subassembly.

3c) The remaining pooled reads are then assembled with IDBA UD to yield subassembled contigs. These subassembled contigs are likely to disambiguate other branches in the scaffold graph because the pooling of *all* reads within the chosen barcodes also draws in reads from flanking regions, due to the long input DNA fragments. This pooling will also draw reads from other input DNA fragments that do *not* cover the local target sequence, which we refer to as off target fragments. These off target fragments will have a low probability of collision with the local target. Nonetheless, to prevent incorrect subassemblies due to off-target reads arising from repeats, we determine a local threshold (based on estimated coverage of the

subassembly target) on the minimum coverage depth required to assemble through a sequence contig. We used the existing short-read assembler IDBA UD to assembled the pooled reads because it was designed for use with highly uneven short-read coverages and also allowed us to specify the minimum support each *k*-mer should have to assemble through a sequence contig. The updated 10X Genomics Chromium platform uses more than an order of magnitude more of droplet partitions than the Gemcode, and should, theoretically, eliminate the need for this additional threshold.

4) The subassembled contigs, which contain large overlaps, together with the initial seed contigs, are passed as reads to the Overlap Layout Consensus Assembler Canu to perform further assembly. Following the methodology used with previous synthetic long read metagenomic assembly approaches[71], we specify a small read error rate to facilitate overlap assembly even in the presence of strain microdiversity. The resulting draft metagenome assembly contains more complete sequence contigs that do not have gaps and resolve shared sequences too difficult to assemble from short-read techniques alone. Repeats that cannot be unambiguously spanned by subassembled contigs remain unresolved by the overlap assembler.

Bibliography

1. Schloss, P. D. & Handelsman, J. Metagenomics for studying unculturable microorganisms: cutting the Gordian knot. *Genome Biol.* 6, 229 (2005).
2. Turnbaugh, P. J. *et al.* An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature* 444, 1027–1031 (2006).
3. Consortium, T. H. M. P. Structure, function and diversity of the healthy human microbiome. *Nature* 486, 207–214 (2012).
4. Taur, Y. *et al.* The effects of intestinal tract bacterial diversity on mortality following allogeneic hematopoietic stem cell transplantation. *Blood* 124, 1174–1182 (2014).
5. Soucy, S. M., Huang, J. & Gogarten, J. P. Horizontal gene transfer: building the web of life. *Nat. Rev. Genet.* 16, 472–482 (2015).
6. Brito, I. L. *et al.* Mobile genes in the human microbiome are structured from global to individual scales. *Nature* 535, 435–439 (2016).
7. Chu, N. D. *et al.* A Mobile Element in mutS Drives Hypermutation in a Marine Vibrio. *MBio* 8, (2017).
8. Kato, N., Yamazoe, K., Han, C.-G. & Ohtsubo, E. New insertion sequence elements in the upstream region of cfiA in imipenem-resistant Bacteroides fragilis strains. *Antimicrob. Agents Chemother.* 47, 979–985 (2003).
9. Glansdorff, N., Charlier, D. & Zafarullah, M. Activation of gene expression by IS2 and IS3. *Cold Spring Harb. Symp. Quant. Biol.* 45 Pt 1, 153–156 (1981).
10. Nagel, M., Reuter, T., Jansen, A., Szekat, C. & Bierbaum, G. Influence of ciprofloxacin and vancomycin on mutation rate and transposition of IS256 in Staphylococcus aureus. *Int. J. Med. Microbiol.* 301, 229–236 (2011).
11. Goussard, S., Sougakoff, W., Mabilat, C., Bauernfeind, A. & Courvalin, P. An IS1-like

element is responsible for high-level synthesis of extended-spectrum beta-lactamase TEM-6 in Enterobacteriaceae. *J. Gen. Microbiol.* 137, 2681–2687 (1991).

12. Peng, Y., Leung, H. C. M., Yiu, S. M. & Chin, F. Y. L. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* 28, 1420–1428 (2012).
13. Namiki, T., Hachiya, T., Tanaka, H. & Sakakibara, Y. MetaVelvet: an extension of Velvet assembler to de novo metagenome assembly from short sequence reads. *Nucleic Acids Res.* 40, e155 (2012).
14. Cleary, B. *et al.* Detection of low-abundance bacterial strains in metagenomic datasets by eigengenome partitioning. *Nat. Biotechnol.* 33, 1053–1060 (2015).
15. Nielsen, H. B. *et al.* Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. *Nat. Biotechnol.* 32, 822–828 (2014).
16. Imelfort, M. *et al.* GroopM: an automated tool for the recovery of population genomes from related metagenomes. *PeerJ* 2, e603 (2014).
17. Wu, Y.-W., Tang, Y.-H., Tringe, S. G., Simmons, B. A. & Singer, S. W. MaxBin: an automated binning method to recover individual genomes from metagenomes using an expectation-maximization algorithm. *Microbiome* 2, 26 (2014).
18. Kang, D. D., Froula, J., Egan, R. & Wang, Z. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ* 3, e1165 (2015).
19. Alneberg, J. *et al.* Binning metagenomic contigs by coverage and composition. *Nat. Methods* 11, 1144–1146 (2014).
20. Chin, C.-S. *et al.* Nonhybrid, finished microbial genome assemblies from long-read SMRT

sequencing data. *Nat. Methods* 10, 563–569 (2013).

21. Loman, N. J., Quick, J. & Simpson, J. T. A complete bacterial genome assembled de novo using only nanopore sequencing data. *Nat. Methods* 12, 733–735 (2015).
22. Koren, S. *et al.* Hybrid error correction and de novo assembly of single-molecule sequencing reads. *Nat. Biotechnol.* 30, 693–700 (2012).
23. Leonard, M. T. *et al.* The methylome of the gut microbiome: disparate Dam methylation patterns in intestinal Bacteroides dorei. *Front. Microbiol.* 5, 361 (2014).
24. He, S. *et al.* Mechanisms of Evolution in High-Consequence Drug Resistance Plasmids. *MBio* 7, (2016).
25. Lange, A. *et al.* Extensive Mobilome-Driven Genome Diversification in Mouse Gut-Associated Bacteroides vulgatus mpk. *Genome Biol. Evol.* 8, 1197–1207 (2016).
26. Kuleshov, V. *et al.* Synthetic long-read sequencing reveals intraspecies diversity in the human microbiome. *Nat. Biotechnol.* advance online publication, (2015).
27. Zheng, G. X. Y. *et al.* Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. *Nat. Biotechnol.* 34, 303–311 (2016).
28. Bishara, A. *et al.* Read clouds uncover variation in complex regions of the human genome. *Genome Res.* 25, 1570–1580 (2015).
29. Peters, B. A. *et al.* Accurate whole-genome sequencing and haplotyping from 10 to 20 human cells. *Nature* 487, 190–195 (2012).
30. Kitzman, J. O. *et al.* Haplotype-resolved genome sequencing of a Gujarati Indian individual. *Nat. Biotechnol.* 29, 59–63 (2011).
31. Amini, S. *et al.* Haplotype-resolved whole-genome sequencing by contiguity-preserving transposition and combinatorial indexing. *Nat. Genet.* 46, 1343–1349 (2014).
32. Koren, S. *et al.* Canu: scalable and accurate long-read assembly via adaptive k-mer

weighting and repeat separation. *Genome Res.* (2017). doi:10.1101/gr.215087.116

33. Klappenbach, J. A., Dunbar, J. M. & Schmidt, T. M. rRNA operon copy number reflects ecological strategies of bacteria. *Appl. Environ. Microbiol.* 66, 1328–1333 (2000).
34. Větrovský, T. & Baldrian, P. The variability of the 16S rRNA gene in bacterial genomes and its consequences for bacterial community analyses. *PLoS One* 8, e57923 (2013).
35. Woese, C. R. & Fox, G. E. Phylogenetic structure of the prokaryotic domain: The primary kingdoms. *Proc. Natl. Acad. Sci. U. S. A.* 74, 5088–5090 (1977).
36. Lagesen, K. *et al.* RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res.* 35, 3100–3108 (2007).
37. Snydman, D. R. *et al.* Lessons learned from the anaerobe survey: historical perspective and review of the most recent data (2005-2007). *Clin. Infect. Dis.* 50 Suppl 1, S26–33 (2010).
38. Desai, M. S. *et al.* A Dietary Fiber-Deprived Gut Microbiota Degrades the Colonic Mucus Barrier and Enhances Pathogen Susceptibility. *Cell* 167, 1339–1353.e21 (2016).
39. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410 (1990).
40. Blakely, G. W. & Sherratt, D. J. Interactions of the site-specific recombinases XerC and XerD with the recombination site dif. *Nucleic Acids Res.* 22, 5613–5620 (1994).
41. Merino, M. *et al.* OXA-24 carbapenemase gene flanked by XerC/XerD-like recombination sites in different plasmids from different Acinetobacter species isolated during a nosocomial outbreak. *Antimicrob. Agents Chemother.* 54, 2724–2727 (2010).
42. Rouquette-Loughlin, C., Dunham, S. A., Kuhn, M., Balthazar, J. T. & Shafer, W. M. The NorM efflux pump of Neisseria gonorrhoeae and Neisseria meningitidis recognizes antimicrobial cationic compounds. *J. Bacteriol.* 185, 1101–1106 (2003).

43. Reeves, A. R., D'Elia, J. N., Frias, J. & Salyers, A. A. A Bacteroides thetaiotaomicron outer membrane protein that is essential for utilization of maltooligosaccharides and starch. *J. Bacteriol.* 178, 823–830 (1996).

44. Erlendsson, L. S., Acheson, R. M., Hederstedt, L. & Le Brun, N. E. Bacillus subtilis ResA is a thiol-disulfide oxidoreductase involved in cytochrome c synthesis. *J. Biol. Chem.* 278, 17852–17858 (2003).

45. Taur, Y. *et al.* Intestinal Domination and the Risk of Bacteremia in Patients Undergoing Allogeneic Hematopoietic Stem Cell Transplantation. *Clin. Infect. Dis.* 55, 905–914 (2012).

46. Weber, D. *et al.* Microbiota Disruption Induced by Early Use of Broad-Spectrum Antibiotics Is an Independent Risk Factor of Outcome after Allogeneic Stem Cell Transplantation. *Biol. Blood Marrow Transplant.* (2017). doi:10.1016/j.bbmt.2017.02.006

47. Mathewson, N. & Reddy, P. The Microbiome and Graft Versus Host Disease. *Curr Stem Cell Rep* 1, 39–47 (2015).

48. Dethlefsen, L. & Relman, D. A. Incomplete recovery and individualized responses of the human distal gut microbiota to repeated antibiotic perturbation. *Proc. Natl. Acad. Sci. U. S. A.* 108 Suppl 1, 4554–4561 (2011).

49. Papadopoulos, D. *et al.* Genomic evolution during a 10,000-generation experiment with bacteria. *Proc. Natl. Acad. Sci. U. S. A.* 96, 3807–3812 (1999).

50. Truong, D. T., Tett, A., Pasolli, E., Huttenhower, C. & Segata, N. Microbial strain-level population structure and genetic diversity from metagenomes. *Genome Res.* (2017). doi:10.1101/gr.216242.116

51. Luo, C. *et al.* ConStrains identifies microbial strains in metagenomic datasets. *Nat. Biotechnol.* advance online publication, (2015).

52. Ward, D. V. *et al.* Metagenomic Sequencing with Strain-Level Resolution Implicates

Uropathogenic E. coli in Necrotizing Enterocolitis and Mortality in Preterm Infants. *Cell Rep.* 14, 2912–2924 (2016).

53. Rinke, C. *et al.* Insights into the phylogeny and coding potential of microbial dark matter. *Nature* 499, 431–437 (2013).
54. Krueger, F. Trim Galore. *A wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files* (2015).
55. Petersen, K. R., Streett, D. A., Gerritsen, A. T., Hunter, S. S. & Settles, M. L. Super Deduper, Fast PCR Duplicate Detection in Fastq Files. in *Proceedings of the 6th ACM Conference on Bioinformatics, Computational Biology and Health Informatics* 491–492 (ACM, 2015).
56. Bankevich, A. *et al.* SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *J. Comput. Biol.* 19, 455–477 (2012-5).
57. Robinson, J. T. *et al.* Integrative genomics viewer. *Nat. Biotechnol.* 29, 24–26 (2011).
58. Ihaka, R. & Gentleman, R. R: A Language for Data Analysis and Graphics. *J. Comput. Graph. Stat.* 5, 299–314 (1996).
59. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis*. (Springer Science & Business Media, 2009).
60. Gu, Z., Gu, L., Eils, R., Schlesner, M. & Brors, B. circlize Implements and enhances circular visualization in R. *Bioinformatics* 30, 2811–2812 (2014).
61. Hunter, J. D. Matplotlib: A 2D Graphics Environment. *Comput. Sci. Eng.* 9, 90–95 (2007).
62. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* 27, 764–770 (2011).
63. Wood, D. & Salzberg, S. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.* 15, R46 (2014).

64. Tatusova, T., Ciufo, S., Fedorov, B., O'Neill, K. & Tolstoy, I. RefSeq microbial genomes database: new representation and annotation strategy. *Nucleic Acids Res.* 42, D553–9 (2014).

65. Benson, D. A. *et al.* GenBank. *Nucleic Acids Res.* 41, D36–42 (2013).

66. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 30, 2068–2069 (2014).

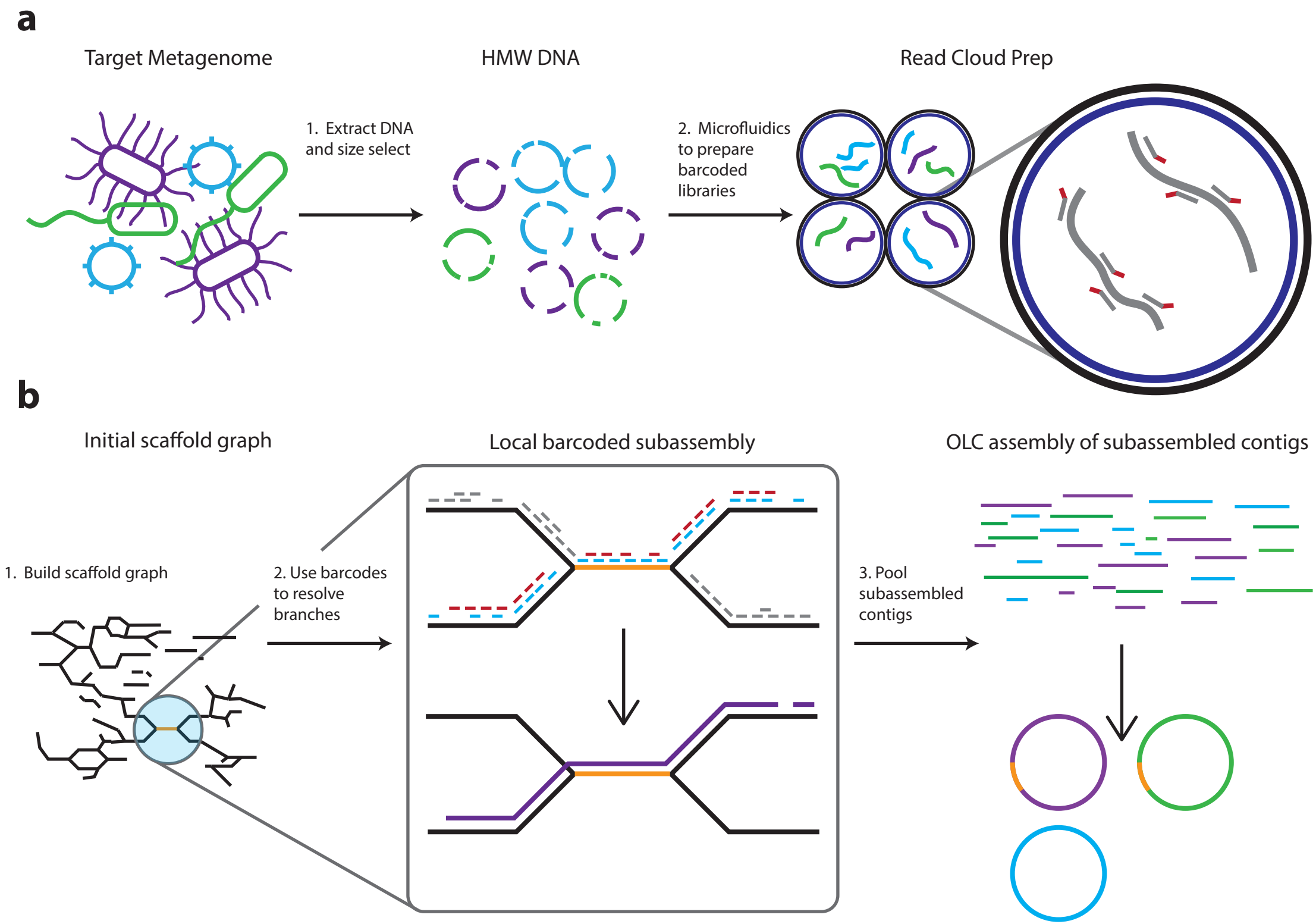
67. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* 25, 1754–1760 (2009).

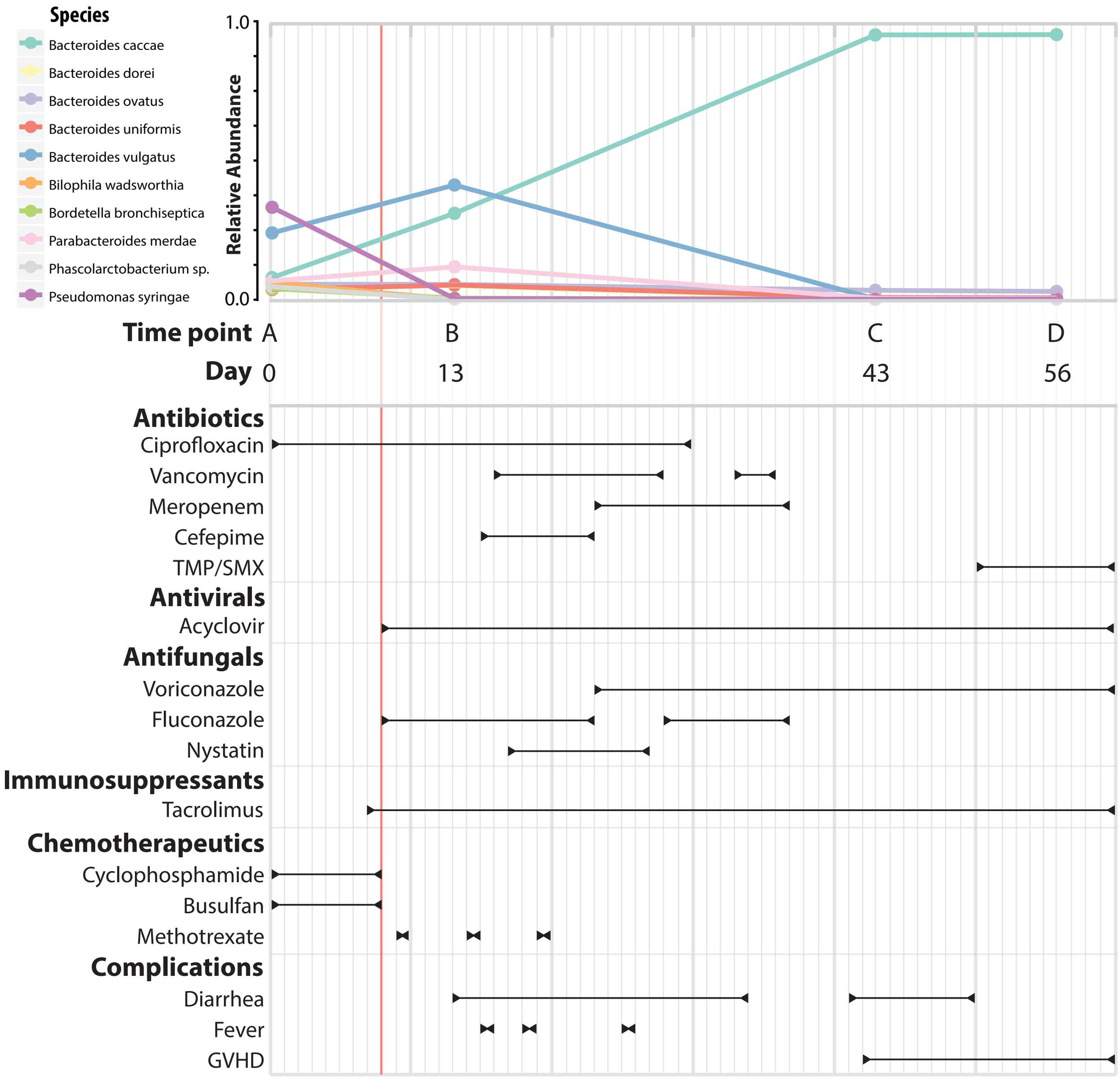
68. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29, 15–21 (2013).

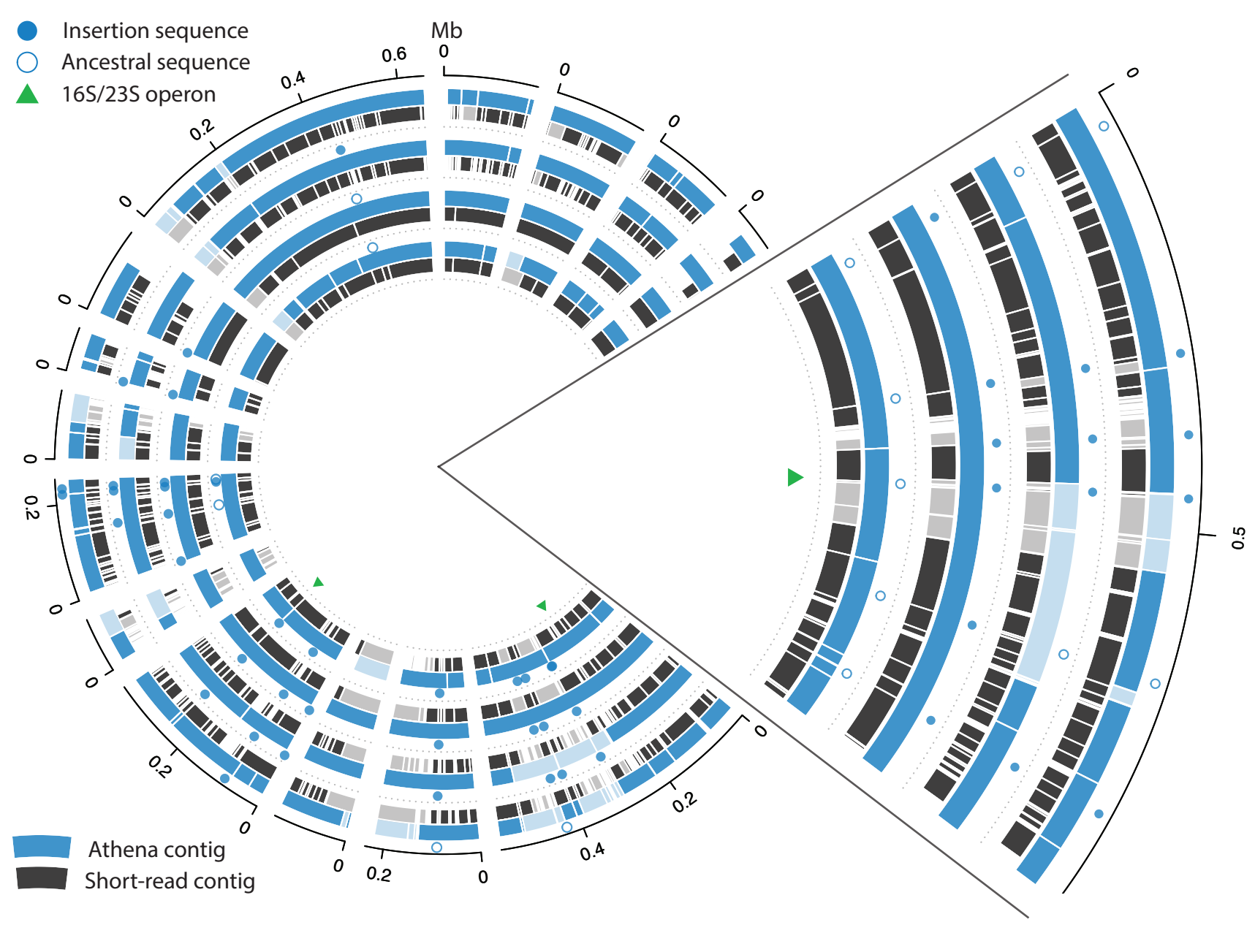
69. Kearse, M. *et al.* Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* 28, 1647–1649 (2012).

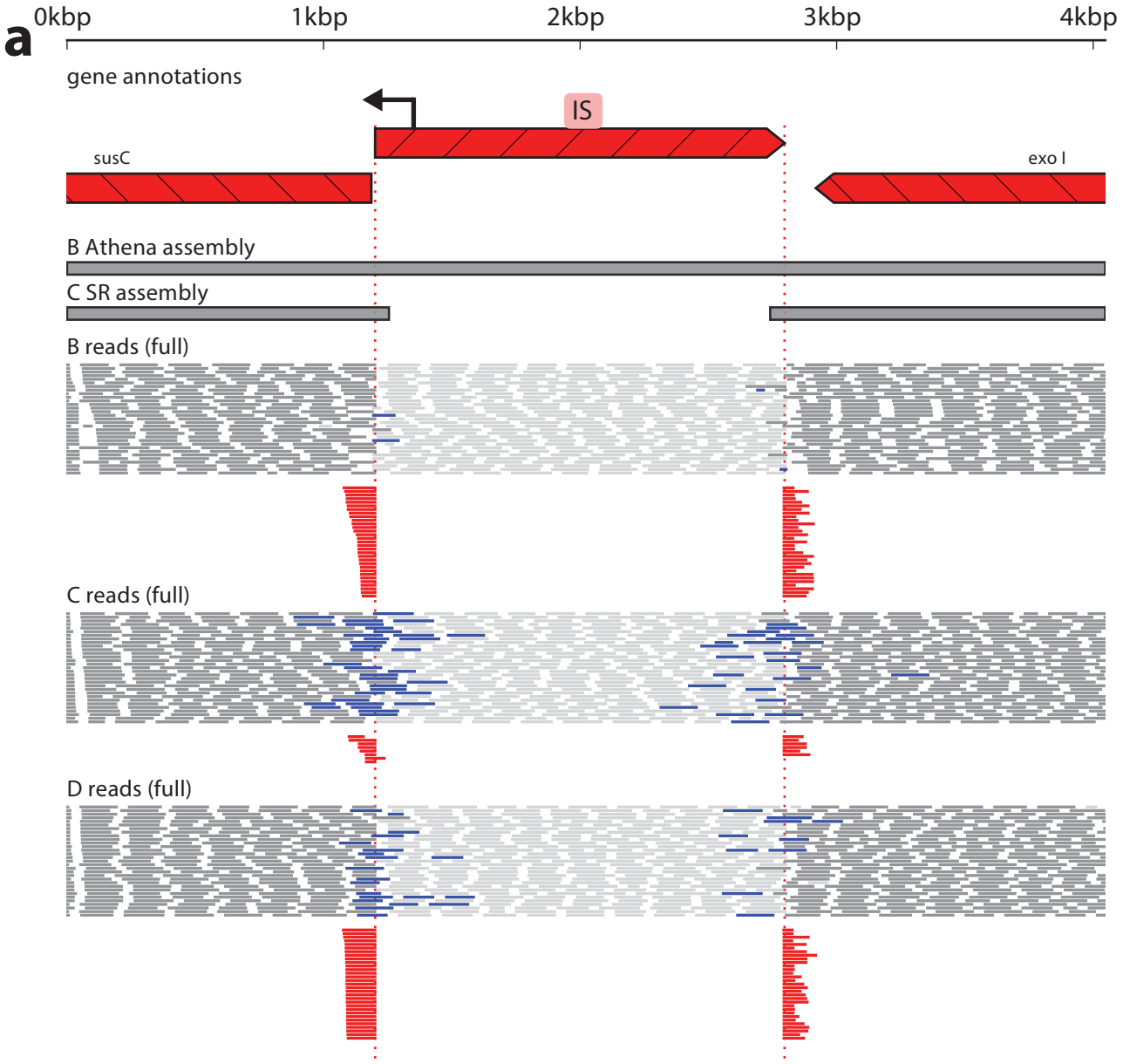
70. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* 25, 1043–1055 (2015).

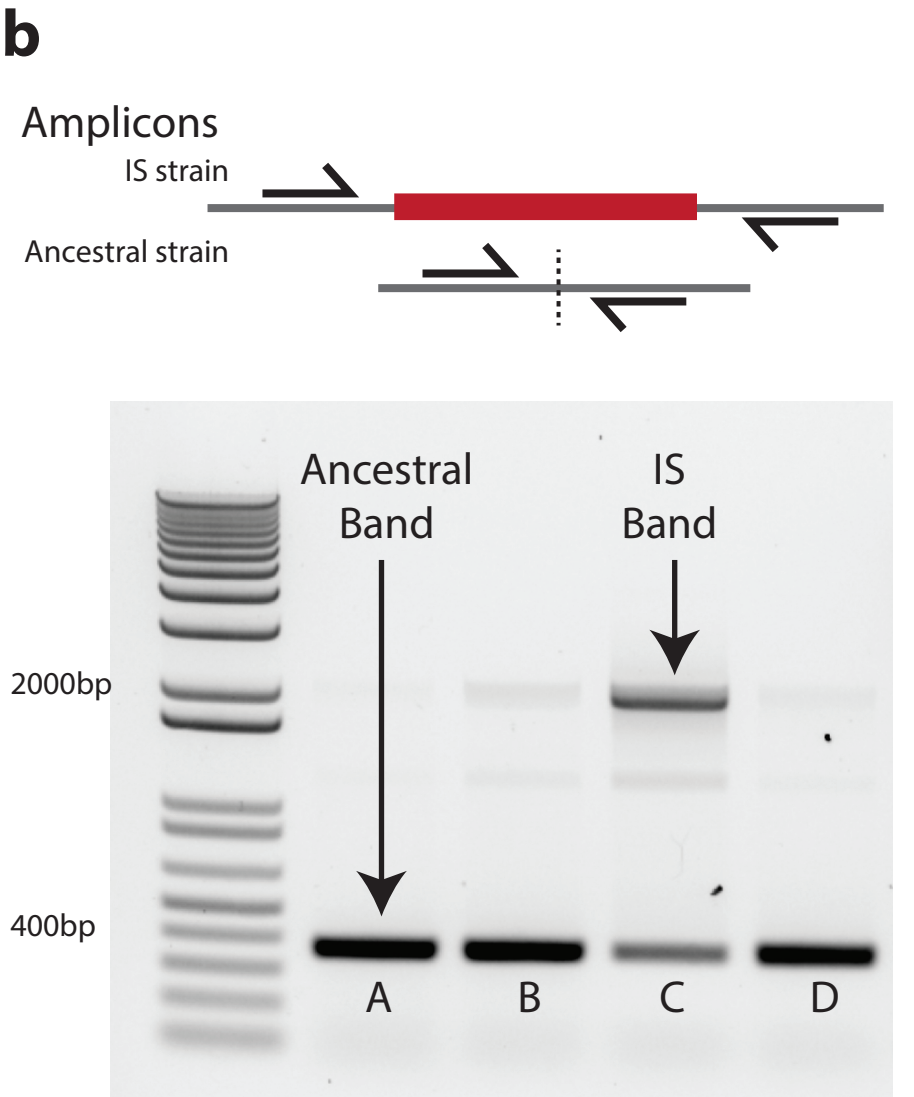
71. Kuleshov, V. *et al.* Synthetic long-read sequencing reveals intraspecies diversity in the human microbiome. *Nat. Biotechnol.* 34, 64–69 (2016).

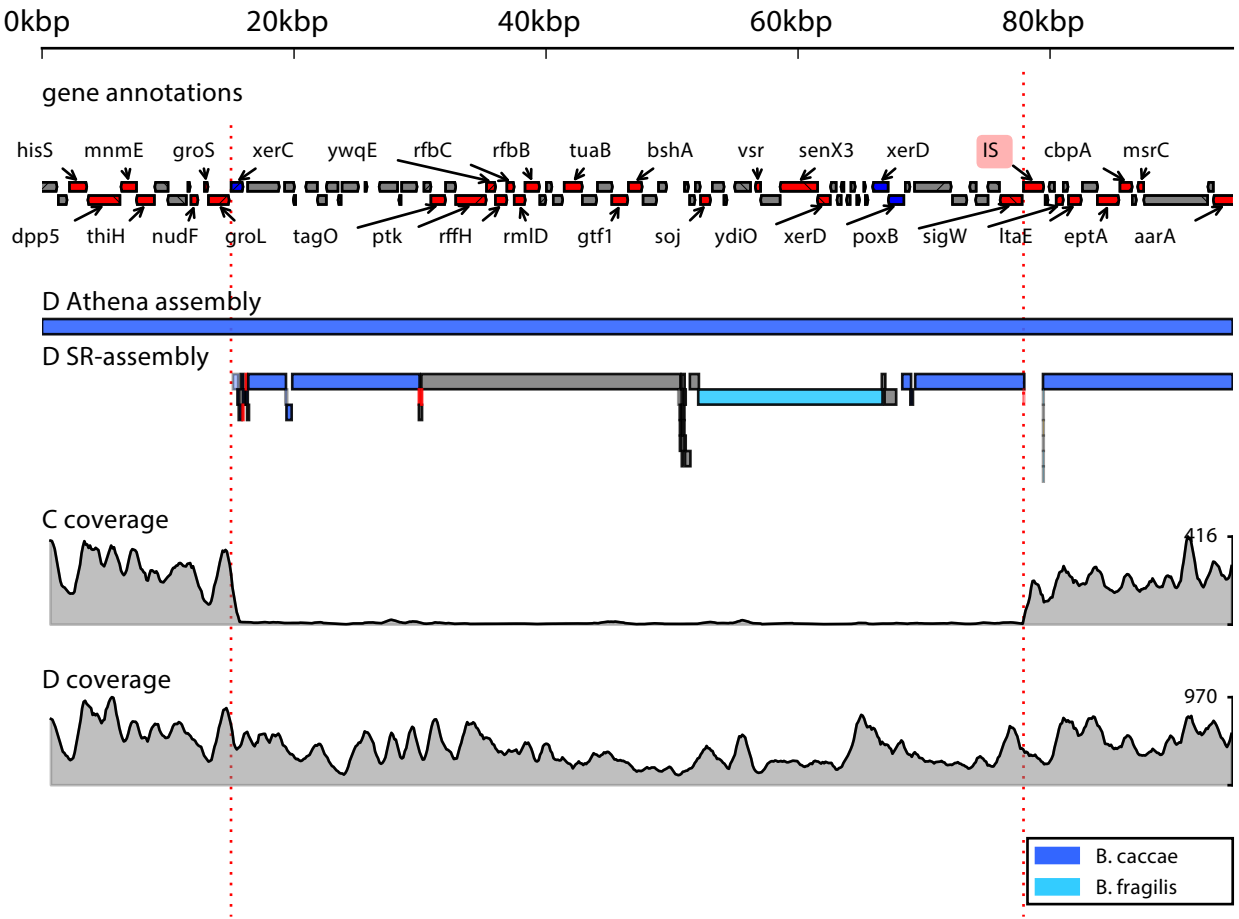

a
Target Metagenome
HMW DNA
Read Cloud Prep
1. Extract DNA and size select
2. Microfluidics to prepare barcoded libraries
b
Initial scaffold graph
Local barcoded subassembly
OLC assembly of subassembled contigs
1. Build scaffold graph
2. Use barcodes to resolve branches
3. Pool subassembled contigs

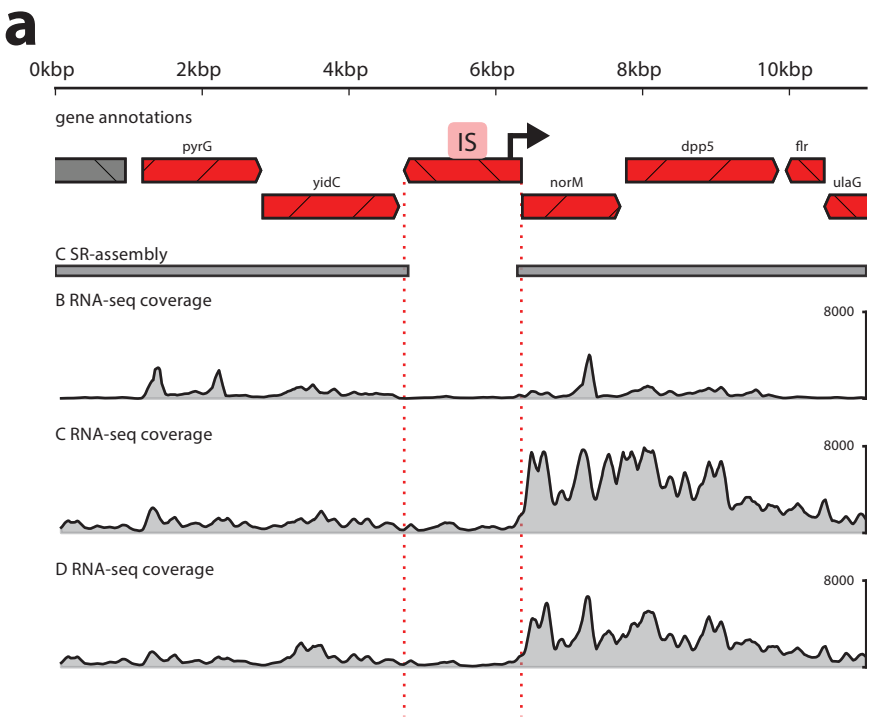

Species
Bacteroides caccae
Bacteroides dorei
Bacteroides ovatus
Bacteroides uniformis
Bacteroides vulgatus
Bilophila wadsworthia
Bordetella bronchiseptica
Parabacteroides merdae
Phascolarctobacterium sp.
Pseudomonas syringae
Relative Abundance
1.0
0.0
Time point A B C D
Day 0 13 43 56
Antibiotics
Ciprofloxacin
Vancomycin
Meropenem
Cefepime
TMP/SMX
Antivirals
Acyclovir
Antifungals
Voriconazole
Fluconazole
Nystatin
Immunosuppressants
Tacrolimus
Chemotherapeutics
Cyclophosphamide
Busulfan
Methotrexate
Complications
Diarrhea
Fever
GVHD

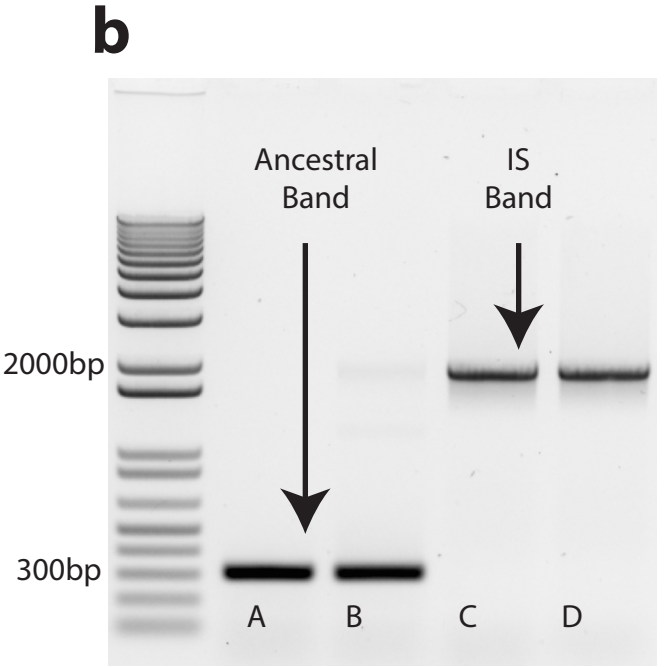

Insertion sequence
Ancestral sequence
16S/23S operon
Mb
0
0.2
0.4
0.6
0.5
Athena contig
Short-read contig


a
0kbp
1kbp
2kbp
3kbp
4kbp
gene annotations
IS
susC
exo I
B Athena assembly
C SR assembly
B reads (full)
C reads (full)
D reads (full)


b
Amplicons
IS strain
Ancestral strain
Ancestral Band
IS Band
2000bp
400bp
A
B
C
D


0kbp
20kbp
40kbp
60kbp
80kbp
gene annotations
hisS
mnmE
groS
xerC
ywqE
rfbC
rfbB
tuaB
bshA
vsr
senX3
xerD
IS
cbpA
msrC
dpp5
thiH
nudF
groL
tagO
ptk
rffH
rmlD
gtf1
soj
ydiO
xerD
poxB
sigW
ltaE
eptA
aarA
D Athena assembly
D SR-assembly
C coverage
416
D coverage
970
B. caccae
B. fragilis


a
0kbp
2kbp
4kbp
6kbp
8kbp
10kbp
gene annotations
pyrG
yidC
IS
norM
dpp5
flr
ulaG
C SR-assembly
B RNA-seq coverage
8000
C RNA-seq coverage
8000
D RNA-seq coverage
8000


b
Ancestral Band
IS Band
2000bp
300bp
A
B
C
D