Published in final edited form as:
Nat Biotechnol. ; 30(7): 701–707. doi:10.1038/nbt.2288.

A Hybrid Approach for the Automated Finishing of Bacterial Genomes

Ali Bashir[1], **Aaron Klammer**[1], **William P. Robins**[4], **Chen-Shan Chin**[1], **Dale Webster**[1], **Ellen Paxinos**[1], **David Hsu**[1], **Meredith Ashby**[1], **Susana Wang**[1], **Paul Peluso**[1], **Robert Sebra**[1], **Jon Sorenson**[1], **James Bullard**[1], **Jackie Yen**[1], **Marie Valdovino**[1], **Emilia Mollova**[1], **Khai Luong**[1], **Steven Lin**[1], **Brianna LaMay**[1], **Amruta Joshi**[1], **Lori Rowe**[2], **Michael Frace**[2], **Cheryl L. Tarr**[2], **Maryann Turnsek**[2], **Brigid M Davis**[3,4,5,6], **Andrew Kasarskis**[1], **John J. Mekalanos**[4], **Matthew K. Waldor**[3,4,5,6], and **Eric E. Schadt**[1,7,*]

[1]Pacific Biosciences, Menlo Park, CA

[2]National Center for Emerging and Zoonotic Infectious Diseases, Centers for Disease Control and Prevention, Atlanta GA 30333

[3]Channing Laboratory, Brigham and Women's Hospital, Boston, MA

[4]Department of Medicine, Harvard Medical School, Boston, MA

[5]Department of Microbiology and Molecular Genetics, Harvard Medical School, Boston, MA

[6]Howard Hughes Medical Institute, Boston, MA

[7]Department of Genetics and Genomic Sciences, Mount Sinai School of Medicine, New York City

Abstract

Dramatic improvements in DNA sequencing technology have revolutionized our ability to characterize most genomic diversity. However, accurate resolution of large structural events has remained challenging due to the comparatively shorter read lengths of second-generation technologies. Emerging third-generation sequencing technologies, which yield markedly increased read length on rapid time scales and for low cost, have the potential to address assembly limitations. Here we combine sequencing data from second- and third-generation DNA sequencing technologies to assemble the two-chromosome genome of a recent Haitian cholera outbreak strain into two nearly finished contigs at > 99.9% accuracy. Complex regions with clinically significant structure were completely resolved. In separate control assemblies on experimental and simulated data for the canonical N16961 reference we obtain 14 and 8 scaffolds greater than 1kb, respectively, correcting several errors in the underlying source data. This work provides a blueprint for the next generation of rapid microbial identification and full-genome assembly.

Introduction

The advent of low-cost, extremely high-throughput second generation sequencing technologies has enabled a dramatic advance in the identification and characterization of microbes. Rather than typing strains based on relatively few selected loci or phenotypes, we can now rapidly sequence bacterial genomes[1–3] thereby enabling detection of subtle inter-strain differences that might not otherwise be recognized[4–8]. However, like most recent

[*]Corresponding author: Eric E. Schadt Pacific Biosciences 1505 Adams Drive Menlo Park, CA 94025 USA Tel: +1 650 521 8025 eschadt@pacificbiosciences.com.

microbial genome projects that have come to rely solely on second generation sequencing technologies, a complete *de novo* assembly of microbial genomes of significant public health concern, like the Haitian outbreak *V. cholerae* strain[5], can not usually be produced. Full genome assembly is confounded by repetitive sequences, which cannot always be accurately placed relative to each other or to the remainder of the genome.

Genome assembly methods have evolved over the years to keep pace with newer sequencing technologies [9–11] by incorporating new computational formulations[12–15]. Today, assembly approaches can be divided into two basic approaches: 1) Overlap-layout-consensus in which sequence reads are aligned to one another to identify overlaps and then leveraging the overlaps to infer layouts of the reads corresponding to the longest contiguous stretches of unique sequence (referred to as a contig)[16, 17]; and 2) the de Bruijn graph approach[18] in which assemblies are represented as special paths through a graph in which pairs of nodes connected by directed edges represent short sequence fragments (reads are interpreted as paths in the graph, and alignments between reads as overlapping paths[13, 19–21]). These approaches have different strengths and weaknesses, depending on the size of the sequencing reads, the number of reads, and complexity of the genome from which the reads were generated. Given significant computational efficiencies, de Bruijn graph based methods have quickly become the preferred computational approach for the *de novo* assembly of next-generation, short read data[19–24], whereas overlap-layout-consensus remains the standard for lower coverage, longer read data.

Despite many computational advances to genome assembly, complete and accurate assembly from second generation short read data remains a major challenge[18, 25–27]. In particular, important aspects of large-scale structural variation and the linear organization of the genome are often missed or incomplete in assemblies based only on second-generation sequence data[18, 26, 2829]. Even in the relatively simple case of bacterial genomes, most genomes based on short-read data remain incomplete[30]; only 26% of those currently deposited into the public domain are completed as of the writing of this paper[31]. Since third-generation sequencing methods generate reads that can span many thousands of bases, they have been expected to facilitate complete genome assemblies and aid in resolution of regions of extended similarity. However, given present the raw accuracy per read from third-generation technologies is significantly lower than from second-generation methods[32], experimental approaches that combine the high accuracy of second-generation data with the longer read lengths of third-generation data could facilitate generation of near-complete and accurate genomes[33–36].

Pacific Biosciences' third-generation SMRT sequencing technology[37] produces sequencing reads by either reading a continuous sequence from the molecular template, or by reading a discontinuous, but linearly ordered, set of reads from the template in a sequencing mode analogous to mate-pair sequencing. Reads from this latter sequencing mode, termed strobe sequencing[38], can span on average more than 6000 bases. In contrast, the continuous reads currently average 3000 bases, although the upper fifth percentile of reads average 9000 bases[2]. The accuracy of the single pass SMRT reads can be improved dramatically using either PacBio circular-consensus sequencing reads[2] or other high-accuracy short reads (e.g. Illumina reads) to error-correct the long reads. Such approaches, however, do not allow facile integration of arbitrary contig data with multiple disparate data types, such as PacBio strobe and continuous long reads, necessitating new algorithms.

Here, we report a hybrid assembly analysis pipeline that combines contigs from an assembly of second-generation sequence data with standard sequencing reads and strobe reads from SMRT sequencing. *De novo* contigs generated from Illumina and Roche 454 sequence data were combined with Pacific Biosciences sequences using a combination of the scaffolding, overlap-layout-consensus and error-correction methods to yield a complete genome of *V. cholerae* isolates from the Haitian cholera outbreak in October 2010[5]. Our hybrid assembly protocol was capable of resolving complex repeat-rich segments of the *V. cholerae* genome, including regions (such as the CTX prophage and flanking areas) where knowledge of chromosome structure has important clinical ramifications. Our approach introduces a blueprint for a new generation of rapid microbial identification and full-genome assembly.

Results

De novo assembly of the Haitian cholera outbreak strain

To implement our hybrid assembly protocol for *de novo* assembly of the genome of *V. cholerae* O1 clonal clinical isolates from Haiti in October 2010 (referred to as H1[5]), we utilized DNA sequence data obtained from PacBio RS sequencers as well as from Illumina GAII and Roche 454 sequencing instruments. The latter data, derived from three independent *V. cholerae* isolates collected in Haiti by the Centers for Disease Control (CDC), had been previously deposited in GenBank (Table 1)[8]. Assemblies of these CDC data (which have GenBank accession numbers AELH00000000.1, AELI00000000.1, and AELJ00000000.1) are comprised of 107, 105, and 93 contigs covering 98.84%, 98.96%, and 98.94% of the genome, respectively, and have N50s of 151kb, 154kb, and 155kb, and maximum contig sizes of 355kb, 344kb, and 500kb, respectively (Table 2). By mapping these contigs to the *V. cholerae* CIRS101 reference genome[39] (the strain we identified as the closest to the outbreak strain), we estimated the overall identity of the CDC assemblies to be 99.99%, likely an underestimate of the assembly accuracy because of the differences that are expected to exist between the genomes of CIRS101 and the Haitian outbreak strain[5]. These contigs comprised a solid genomic foundation for augmentation, using our hybrid assembly strategy and reads generated on the PacBio RS, into a fully assembled genome.

Two types of data were generated from the PacBio RS and prototype instruments: 1) continuous long read data (reads ~1250 bp); and 2) strobe read data representing intermittent reads (~500bp) of contiguous DNA sequence flanking long stretches of template sequence that are synthesized but not directly observed[38, 40]. The strobe reads generated on H1 averaged 6180 bp with a mode exceeding 6600 bp (Supplementary Figure 1). We initially generated 19× and 26× of physical coverage for long read and strobe read data, respectively (Table 1), which when combined with the CDC data yielded the initial two chromosome scaffolds. Subsequently, we generated 138X coverage of long read data using the latest chemistry (C2) and operating characteristics of the PacBio RS (Supplementary Figure 2) to complete the genome assembly as discussed below.

To carry out the hybrid assembly procedure, we first generated a consensus CDC contig set. Given the clonal nature of the CDC isolates (Supplementary Results), contigs from the minimal CDC assembly that were inconsistent with the remaining two isolates were broken up. If the split resulted in a subcontig less than 1kb in length, the subcontig was eliminated. The resulting 97 contigs in this set, along with 68,565 single molecule reads from the PacBio RS with an average accuracy of 85% (Supplementary Figure 2), were input into our hybrid assembly pipeline (Supplementary Figure 3). The pipeline starts with five inputs: the CDC contigs, the raw 454 reads, the raw Illumina reads, and the long and strobe PacBio reads. PacBio long reads and strobe reads were used to scaffold the CDC contigs with the AHA scaffolding algorithm (Methods). The PacBio long reads were also corrected with the 454 and Illumina reads using a previously described method[2], and these corrected reads were used to either error-correct or fill in gaps in the AHA scaffolds at various points of the pipeline. Finally, the resulting scaffolds were re-sequenced using the PacBio long reads, and consensus was called using these reads, ensuring that the final sequence came from a single clonal source. The computational performance of the assembly pipeline is given in Supplementary Table 1.

The resulting hybrid assembly was comprised of two circular contigs, each representing one of the chromosomes of the *V. cholerae* genome (Figure 1). The overall accuracy of the consensus sequence for this assembly was 99.99%, determined by comparing the consensus sequence to the CIRS101 genome (Table 2). Sanger sequencing was also performed after assembly to confirm linking of 47 contigs (see Supplementary Methods, also Figure 1); these analyses also indicated that extensive confidence in the assembly pipeline is warranted. We note that data from all three platforms contributed to the successful assembly. Without the long reads generated from the PacBio platform, the hybrid Illumina/454 assembly yielded 93 contigs (Table 2). Furthermore, the strobe reads with 6.2kb spans was critical for producing the two-contig assembly; without these reads, a hybrid assembly derived from only the Illumina, 454, and long reads yielded 45 contigs. It is important to note that the initial assembly process prior to the validation and refinement of complex regions in the cholera genome like CTX using the C2 data was completely automated and resulted in a two-scaffold assembly. A manual inspection step was performed that involving looking at the scaffold graph from the first iteration of AHA and identifying CDC contigs that were involved in complex repeat tangles. The motivation for this manual step was a concern that in complex regions like CTX and superintegron the original CDC contigs could have been misassembled. For example, even if repeats in these regions were properly duplicated in AHA, the original CDC contigs could represent consensus repeats; a given repeat instance may exist as a truncated copy in the underlying genome. To address this issue, these two regions were reassembled by using long reads with less dependence on the CDC contigs. Specifically, high-quality anchors were identified adjacent to each region. We then used the rough scaffold layout as a guide and gap-filled this initial scaffold with long reads. While the assembly process did involve this manual examination to resolve tangles in the graph representing the assembly, we note that the two-contig hybrid assembly was achieved without the need for additional experimental steps, such as the construction of fosmid libraries or targeted PCR followed by Sanger sequencing, that are currently routinely used to close gaps or assess structures of specific regions to finish genome assembly. We further note that the consensus accuracy achieved using the Illumina, 454, and PacBio data combined was higher than using a subset of these data (Supplementary Table 2).

Given the CDC isolates were collected independently of our H1 isolate, we sought to ensure the PacBio sequences for H1 were structurally consistent with the CDC contigs and that the CDC contigs were themselves self-consistent. Because the base calls in our final assembly are made from the Illumina and PacBio data we generated on the H1 strain, only structural inconsistencies between the CDC isolates and our H1 isolate have the potential to grossly mislead the assembly process. To assess this we carried out a series of detailed comparisons between the long read PacBio data and the CDC contigs used in the assembly process described above, with and without synthetic breakpoints introduced into these contigs (the synthetic breakpoints served as a positive control to highlight our pipeline's ability to detect misassemblies). For each position in the different reference sequences, we determined whether there was at least one PacBio long read of high quality that was structurally consistent with a 200 base pair window around the test position. If there were structural inconsistencies between the CDC isolates and H1, then we would expect those regions that are structurally inconsistent to not be supported by any PacBio reads without gaps. In fact, we found that for each of the CDC isolates across every base position greater than 200bp away from the ends of the contigs (for a fuller analysis of this consistency mapping strategy including edge effects, see Supplementary Results) that there existed a high confidence PacBio read that was structurally consistent with the reference sequence. In contrast, for the set of CDC contigs for which synthetic breakpoints were introduced, we failed to identify any PacBio reads that were structurally consistent with the synthetic break point regions. We repeated this analysis with larger (500 base pair) overhang lengths. The only structurally inconsistent regions detected in any of the isolate assemblies appeared to be regions of repeat compression or expansion, suggesting pathologies in the underlying assembly algorithm used to assemble the data as opposed to real differences between strains (See Supplementary Results). These data combined confirm our initial analysis[5] and the analysis of other groups[40] that the Haitian outbreak strain was the same across these isolates.

Resolving complex chromosome regions with long read data

Three components of the *V. cholerae* genome that consist of complex repeat structures were particularly challenging to assemble and highlight the utility of long reads to complete even relatively small, moderately complex bacterial genomes. These sequences – consisting of rRNA operons, the CTX/RS1/TLC prophage regions, and the superintegron (SI) on *V. cholerae* chromosome II – corresponded to numerous relatively short contigs in the CDC data set (Figure 1); determination of their relative and absolute positions within the genome was not possible without the long and strobe read sequence data, given the complex repeat structures within these regions that required long reads in order to unambiguously resolve them. Critical knowledge can be obtained by achieving accurate assembly of such complex regions. For example, in *V. cholerae*, the linear structure of one of these complex repeat rich regions (CTX/RS1/TLC prophage region) is known to be essential for this pathogen to disseminate *ctxAB*, the genes encoding its key virulence factor, cholera toxin[41].

Placing the ribosomal RNA operon repeats—The ribosomal RNA (rRNA) operon, encoding 16S, 23S, and 5S rRNA, is typically greater than 5kb in size and occurs seven times in the outbreak strain genome, with an average sequence identity between the repeats ranging from 98.04% to 99.94% and structural differences greater than 500bp). Given the high degree of sequence identity among these repeats, placing the different repeat elements without spanning them is difficult or impossible; in fact, rRNA operons were present within 7 of the 45 gaps in the Illumina/454 assembly. However, as we had at least three strobe reads spanning each of the 7 repeats, we were able to unambiguously resolve links between CDC contigs that flank rRNA operons (Figure 2). Using the latest recently described chemistry for the PacBio RS[2], we generated additional long reads with an average read length of 2800bp (Figure 2 panel D), which provided complete structural information for each repeat instance. Figure 2 (panel E) indicates the mapping of CDC contigs to each of the assembled rRNA operon repeats in H1. We highlight an insertion in one of the rRNA subunits compared to the others (the RNA operon linking contigs 80 and 59 has an additional contig 4, indicated by the circled blue region), demonstrating that variation between the repeats can be localized to a specific repeat locus.

CTX Prophage/TLC Region—For the seventh pandemic of cholera that began in 1961, the El Tor O1 strains that are causing this ongoing pandemic can be distinguished in part by differences among their arrays of CTX prophages and RS1 and TLC satellite prophages, which lie near the terminus of the large chromosome[7]. However, the repetitive sequences among these arrays can confound genomic assembly. In particular, the presence of *rstR, rstA*, and *rstB* within both CTX prophages and RS1 can make it difficult to determine their relative positions and copy numbers. Subtle differences among TLC elements in a single strain can also complicate genome assembly. The CDC assembly of Illumina/454 sequences covering the CTX/TLC region of the Haitian cholera outbreak strain consisted of 10 contigs that could not be unambiguously ordered nor could the number of repeats for each contig be determined. Notably, by combining these contigs with strobe and long reads, and then validating and refining the resulting scaffolds with C2 data, we were able to fully assemble this region. Long reads were particularly useful for clarification of tandem repeat structures in these instances (Figure 3A–C). Notably, the position of the single CTX prophage relative to the single RS1 (RS1 upstream of CTX) suggests that the Haitian strain should not be able to produce CTX virions by using the chromosome initiated rolling circle type mechanism[41], and thus cannot disseminate the genes encoding cholera toxin to other *V. cholerae* isolates. A similar arrangement of CTX/RS1/elements is seen in CIRS101, a 2002 *V. cholerae* from Bangladesh[39].

Our complete assembly indicates that the Haitian *V. cholerae* outbreak strain (H1) contains two copies of the TLC element adjacent to its CTX prophage array (Figure 3). Such tandem arrays have been detected in numerous *V. cholerae* isolates from the 7th pandemic of cholera[42, 43], including N16961, the first strain for which genomic sequence became available. However, the structure of this region was not previously known for the outbreak strain or the closely related strain CIRS101, as sequence contigs generated from sequencing results were fragmented. Our sequence assembly also confirms some notable features of the H1 TLC region, which may be useful as benchmarks for strain comparisons. In particular, there is a transposase (contig 11) adjacent to the TLC repeats in H1 that is absent from this site in N16961 and many other 7th pandemic strains. To date, this transposase (which is more typically found within the Vibrio SXT element) has predominantly been detected adjacent to TLC in recent *V. cholerae* isolates that harbor the classical ctxB variant of the CTX prophage, including CIRS101 and a recently sequenced isolate from Nepal, which is hypothesized to be the source of the Haitian outbreak strain[6].

The assembly of CTX/RS1/TLC region was validated in two ways. First we generated and sequenced long range PCR products over this region. Two sets of PCR primers were designed, each beginning from an anchoring point outside of repeat regions near the TLC/CTX region and ending at a position spanning CTX or TLC, respectively. Long read sequencing of the resultant PCR products (Figure 3F) enabled validation of the gene/contig ordering. For the CTX region, since the PCR fragment size was smaller (5.6kb), many reads spanned the entire region, giving unambiguous confirmation of the product. In contrast, only fragments from within TLC PCR products were sequenced; still, the resulting data was concordant with the assembly. Secondly, Southern blotting also confirmed that the H1 isolate, like N16961, contains a tandem duplication of the TLC element, as well as revealing a RFLP consistent with the nearby transposon insertion (contig 11) in H1 (Supplementary Figure 4). As a control we sequenced an amplicon containing the tandemly duplicated TLC element from the reference strain N16961 which was consistent with the canonical N16961 reference (Supplementary Figure 5).

Superintegron Region—The ~125Kb superintegron in chromosome II is composed of a large number of unrelated cassettes separated by repeated and related DNA elements[44–46]. The low complexity of this region makes it difficult to resolve using second-generation sequencing assembly tools, and accounts for the fragmentation observed in the CDC contigs associated with superintegron sequences, which were on average significantly shorter than those covering the remainder of the genome (2.2Kb vs 69Kb) (Figure 4D). However, strobe reads enabled ordering of unique contigs, and continuous reads allowed for fill-in of the entire interval between such contigs. The significantly larger C2 read length (Figure 4A) compared to the maximally repeated contig size allows resolution and duplication of multiple ordered contigs within a single read. Collectively, these data sets allowed for comprehensive assembly of the superintegron, thereby facilitating inter-strain comparisons of this highly variable region, which may assist in gain-of-fitness within an environmental niche or possibly play a role in pathogenesis[44].

Validating the assembly pipeline on N16961 *V. cholerae*

To validate whether the quality of the assembly we achieved on the Haitian outbreak strain using our hybrid assembly approach can be achieved beyond this specific example, and to assess its validity for Cholera assemblies, we applied it to the N16961 strain of cholera. The genome sequence of N16961, a strain isolated in Bangladesh in 1971, has served as the canonical reference genome for seventh pandemic *V. cholerae* strains.

We combined PacBio data with 454 data generated on the N16961 control strain (Supplementary Table 3). The 454 data were assembled *de novo* into contigs using Newbler (See Supplementary Methods), then these contigs and the PacBio data were input into our hybrid assembly procedure. The assembly in this case was comprised of 14 scaffolds (18 contigs) containing 6,744 uncalled bases, with 8 scaffolds covering 99% of the genome, (Supplementary Table 4). We next sought to assess the accuracy of our automated approach for H1 using N16961. Because our assembly of the H1 strain used local reassembly (manually extracting a subset of CDC contigs) to assemble the CTX, Superintegron and 5kb rRNA operons, we excluded the corresponding regions in N16961 from our accuracy calculations for our N16961 assembly. We aligned the remaining regions of the canonical reference to our assembly yielding an accuracy of 99.98%.

The accuracy improved to 99.99 after error-correcting with simulated 36 bp Illumina reads (at 1% error) and resequencing with the PacBio long read data (Supplementary Table 4). This calculation excludes errors from a 353 base pair indel contained within a tandem repeat that is confirmed by the raw PacBio data (Supplementary Table 5 and Supplementary Figure 6). Most uncalled bases (gaps) in the scaffolds resulted from the strobe read data, although in some cases the gaps were supported by low accuracy sequence data (Supplementary Tables 6 and 7). There were many instances in which our scaffolding and error-correction process corrected small nucleotide variation errors (mainly indels) introduced during the assembly of the 454 data into contigs. In one instance the initial 454 assembly had a large insertion (several hundred bases) as determined by comparison to the existing, complete N16961 reference. The largest single correction resulted in adjustment of a 440 bp deletion in the 454 contigs relative to the N16961 reference (Supplementary Results).

To further validate our H1 assembly process, we simulated starting contigs, fragmenting the known N16961 reference genome in regions syntenic to the H1 starting contigs and then combined these data with PacBio long read and strobe read data generated on this strain (Supplementary Methods and Supplementary Table 3). After layering in the PacBio long and strobe read data for N16961, the hybrid assembly procedure resulted in an assembly comprised of 11 scaffolds and an accuracy of 99.99% (Supplementary Table 4).

Discussion

Changes in bacterial phenotypes are not only induced by small nucleotide variations, but also by larger structural variations. The relative positions of genes, regulatory regions, and other genome components can affect key biological processes, including pathogenicity and drug resistance. Only by assembling genomes in a *de novo* fashion, without reference genomes to guide the assembly process, can we objectively characterize the complete complement of genetic variation between strains, not only explaining relatedness among the strains, but key functional differences as well.

Here we have demonstrated a novel hybrid *de novo* assembly procedure in which high-accuracy short read data from second-generation sequencing technologies were combined with long read PacBio data to completely assemble the genome sequence of the Haitian cholera outbreak strain. Our method is applicable for augmenting extant assemblies from diverse sources, e.g. Sanger and short read sequencing. We found that long reads accurately resolved the structure of complex regions in the *V. cholerae* genome, like the CTX prophage and superintegron regions, as well as properly placed large repeat elements like the ribosomal RNA operon repeat. Resolution of these repeat-rich regions was a direct result of the coverage of these regions by long reads. Characterizing the structure of regions harboring repeats with high sequence similarity requires reads that unambiguously anchor on either side of the repeat element. In the case of the CTX region, the position of the CTX prophage relative to other components like RS1 is critical information since the position defines whether CTX virions will be produced or not[41].

The genome sequence of N16961, a strain isolated in Bangladesh in 1971, has served as the canonical reference genome for seventh pandemic *V. cholerae* strains. Since the onset of the seventh pandemic of cholera in 1961, the El Tor O1 strains that are causing this ongoing pandemic have continued to evolve[7]. Notably, our complete genome sequence of the Haitian cholera outbreak strain differs in interesting ways from N16961 and it more closely resembles strains isolated in Asia during the last decade[2]; furthermore it is nearly identical to the 2010 Nepalese cholera outbreak strain that was recently sequenced[6]. A phylogenetic analysis of SNP data derived from the sequencing of a number of isolates from the 2010 Nepalese cholera outbreak demonstrated that the Haitian outbreak strain was much more closely related to the most recent Nepalese outbreak strain than to N16961[6]. Similarly, we found that at the structural level the Nepalese strain was more closely related to the Haitian outbreak strain than to N16961 (Supplementary Figure 9). Collectively, these observations suggest that the complete sequence of the Haitian cholera outbreak strain may serve as an appropriate reference for more recent seventh pandemic *V. cholerae* isolates.

Our hybrid *de novo* assembly protocol should be applicable for completing the genomes of currently incomplete bacterial genomes in GenBank as well as for generating complete genomes of bacteria yet to be sequenced. While the current implementation of our assembler can only easily handle genomes on the order of 150 Mb, given the random access memory requirements for the current implementation of our assembly algorithm for large genomes (> 1 gigabases) is beyond what resides in a typical compute node (Supplementary Table 1), we expect future improvements will help scale the approach to mammalian-size genomes.

Third generation technologies are particularly relevant for hybrid *de novo* genome assembly problems when a reasonable number of gaps remaining from assemblies based on short read data are within the differences between the expected length distribution of the third generation platform versus the initial second generation platform (Supplementary Figure 10). Complete genomes are useful for bacterial identification and typing, annotation, and deciphering genomic structure. Furthermore, as illustrated here by the structure of the CTX prophage region, knowledge of the correct structure of repeat regions can have important functional consequences. The information provided by complete or near complete genomes for understanding newly emerged pathogens, like Shiga toxin producing *E. coli* 104:H4[2], is clearly far superior to low-resolution techniques, like serotyping or pulsed-field gel electrophoresis (PFGE). Compared to techniques like PFGE, whole genome sequencing of pathogens yields far more granular, high resolution, and complete understanding, down to the single nucleotide level, thereby enhancing our ability to diagnose, track, and ultimately prevent the spread of disease[2].

Methods

Hybrid Assembler, *AHA*

The AHA scaffolding algorithm is the heart of the hybrid assembly pipeline, and was applied twice to scaffold the CDC contigs. AHA first builds a scaffold by aligning reads to the contigs, using reads that span multiple contigs as links to build a scaffold graph. AHA resolves (untangles) complicated structures in the scaffold graph to return a set of linearized scaffolds with “Ns” indicating the expected span distance between contigs (Supplementary Figure 11). This process is repeated within AHA, replacing the initial contigs with the linear scaffolds output from the untangling process and reducing the thresholds for links used in the scaffold graph. Gaps (denoted as “Ns”) between contigs were then replaced with consensus from overhanging reads. The scaffold output from the long read scaffolding was used as input to the second application of the AHA scaffolding algorithm, which was carried out in a similar fashion as the first phase, but instead using the strobe reads to link contigs together (Supplementary Figure 3).

Selection of linking reads—PacBio reads were aligned to the input contigs using BLASR (http://www.pacbiodevnet.com/). BLASR parameters used for the study were: “-minMatch 6 -minFrac 0.1 -minPctIdentity 60 -bestn 10”. From each read we filtered the alignments by quality (see *Iteration Schedule* below) and screened the alignments for uniqueness. An alignment was considered unique if it did not overlap (in called bases on the read) another alignment on the same read by at least 5%. If two alignments overlapped and the difference between the BLASR scores was less than 100, the alignments were considered ambiguous and ignored. In the cases where the score difference was greater than 100 the alignment was considered unambiguous, and the placement with the better score was selected.

Two types of spans were passed into Bambus[47]: 1) spans from long reads spanning multiple contigs, and 2) spans from strobe advances. For long reads with multiple unambiguous alignments, as described above, we computed the pairwise spans between each alignment pair. The mean span was given by the number of called bases between the last mapped base of the current alignment to the first mapped base of the next alignment, with the minimum and maximum span by min(.2 * μ, 50), where μ is the mean span. For all aligned strobe subread pairs we enforced a span distribution of 4000–9000bp with a mean span length of 6500 bp. Any alignments outside this distribution would not contribute to acceptable edges.

Repeat Rescoring—For each alignment we rescored the alignment by the fraction of the read within a repeat. Repeats were determined by alignment of the input contigs to each other. The alignment was performed using the tool nucmer in the Mummer package[48] with the parameters “—nosimplify -- maxmatch”. Only regions of identity greater than 90% were considered for rescoring alignments. The rescoring simply adjusts the score and alignment length of each alignment by a factor equal to the fraction of the alignment contained outside the repeat sequence.

Layout/linearization of scaffolds and repeat duplication—The linkage information was passed to Bambus along with a redundancy threshold. Each of the libraries was given equal weight and overlaps between contigs were not permitted. The Bambus graph was then converted into a GraphML document (details of which are given below) including all edges that were not rejected by Bambus, even those labeled as “UNSEEN.” The GraphML document was then used as input to a custom untangling algorithm.

In the first step of the untangling procedure, densely connected subgraphs that can be turned into linear paths were identified using the ordering provided by the Bambus layout (when available), as shown in Supplementary Figures 11A and 11B. Each node in the subgraph between the source and sink nodes must be consistent with the span information of its spanning edges. Specifically, each node must be smaller in size than the maximal allowable span of edges that span it. If there existed edges from nodes within the subgraph to nodes *outside* the subgraph, each of these nodes were duplicated so that they could be represented as repeats and take part in linking other contigs. All edges to outside nodes were then removed from the original node (internal to the subgraph) and re-connected to the duplicated node (Supplementary Figure 11C). The simplified graph was then evaluated for more complex structures in which absent edges were inferred from spanning edges and Bambus layout information (Supplementary Figure 11D). For each of the inferred edges the new edges span constraints were created by adjusting spanning edges by the size of the contig and its span distance to its nearest common ancestor, as show in Supplementary Figure 11D.

GraphML file format—GraphML is an XML representation for arbitrary graphs that we used to represent our scaffolds at various stages of the AHA pipeline. In the graphs, contigs are represented as nodes and sequences that link contigs (either strobe or long reads) are represented as edges. Each node and edge can have attributes describing its features. The GraphML files can be read into python objects using the networkx (http://networkx.lanl.gov/) python package.

Iteration schedule—At each iteration AHA realigns reads back to the scaffold in order to further refine the assembly, an approach motivated by the observation that as the assembly improves additional linking reads can be found, especially at the junctions of contigs. Three parameters are evaluated at each iteration: 1) an alignment score (per subread mapping), 2) edge redundancy (minimum number of spanning reads in order to consider an edge valid), and 3) minimum length of alignment. Though different parameters can be used at each stage, for consistency, the same parameter set was used for both the long read and strobe read AHA stages

Error-correction of the PacBio RS sequence data

To complete the hybrid assembly process we recruited high-coverage Illumina short reads to correct the low-coverage PacBio sequences used to fill-in gaps. Short reads were aligned to the PacBio-only regions using BLASR[2]. Regions in which short-reads could be mapped were corrected by taking the consensus of short-reads overlapping the region. We note that while in this case Illumina short read data were used for error correcting the PacBio RS long reads, circular consensus sequence data could be generated as well using the PacBio RS, providing a way to *de novo* assemble genomes using a single technology[2]. In this present case, the existence of short-read Illumina sequence data obviated the need to generate additional PacBio RS data for error correction.

Testing the Allora assembly pipeline on simulated data

We found *de novo* assembly of PacBio reads at current error rates of approximately 16% to be challenging. As such we sought to understand how the ability to *de novo* assemble PacBio-like reads varied as a function of error-rate. We simulated 20× reads from the MJ1236 cholera genome at varying error rates using an error profile similar to that of PacBio reads and assembled these reads using Allora[2], the PacBio *de novo* assembly pipeline. The Allora assembler is based on the AMOS software package, and uses a traditional overlap-layout-consensus approach to assembly. Errors were distributed across the simulated reads in the following proportions: insertion 50%, deletion 40%, substitution 10% (thus, under this model a sequence with 10% total error rate would have 5% insertion, 4% deletion and 1% substitution errors). Performance statistics for each assembly are shown in Supplementary Table 9.

Testing the AHA pipeline on the N16961 cholera strain

De novo assembly of the N16961 strain was carried out as described above for simulated data, but substituting in 454 data generated on N16961 in place of the simulated contigs.

Sample Vibrio_N16961 was sequenced using a Roche 454 GS-FLX+ sequencer using Titanium FLX+ chemistry and the standard Roche shotgun library protocol (454 Life Sciences/Roche Applied Science). A total of 366,623 thousand reads (169 Mb; median read length 527 bases) were assembled using the Newbler assembler V2.6 (release 20110517_1502). The resulting assembly comprised 68 large contigs (>1kb base cutoff) and 120 total contigs (N_{50} = 197,662 bases). This resulted in 42× average coverage of a 3.96 Mb draft genome with 99.90% Q_{40} bases (Supplementary Tables 3 and 4).

Assembly of rRNA operons, CTX/TLC, and Superintegron

In order to ensure accuracy, several additional steps were taken in the assembly of the complex rRNA repeats, CTX/TLC region, and superintegron region. First, it was observed that the CDC contigs constituting the rRNA operon were not identical for every occurrence of the repeat in the genome. Since the AHA hybrid assembly algorithm relies on the input contigs being correct, we decided to exclude these contigs from our initial analysis, allowing them to participate later on in error correction of the assembly. The occurrences were scaffolded via strobe sequencing and the gaps were subsequently filled with PacBio only sequences so as not to bias the repeats to fit the CDC consensus contigs.

Second, two highly repetitive regions (CTX/TLC and superintegron) were locally reassembled (without the entire contig set). Most of the repeats in these regions were shorter than 1kb. Thus, to simplify the scaffolding process all contigs less than 1kb were excluded from these local assemblies and the resulting gaps were filled in using PacBio long reads. The predicted assembly in the CTX region was validated by PCR and Southern blotting (Figure 3 and Supplementary Figure 5). Overlapping PCR products were designed spanning the entire predicted assembly. The products were sequenced using PacBio long reads.

Filling gaps between CDC Contigs

The linear scaffold of CDC contigs may have small (<10kb) gaps corresponding to regions of DNA for which no corresponding input contig exists. For each of these regions, we “gap filled” between the CDC contig pair. All 454 error-corrected PacBio continuous long reads overhanging the gap were extracted. For large gaps (>2kb) strobe subreads overhanging the junction were also recruited. Overhanging strobes were defined as strobes with a single mapped subread whose unmapped subreads could reside in or overlap a gap (i.e. the distance of the mapped subread to the gap was within the minimum and maximum span distribution and the subread was in the correct orientation).

These reads were then aligned to each other as well as the two contigs flanking the gap. The resulting alignments were used to produce an overlap graph in which reads (or contigs) are nodes and edges are overlaps between the reads (or contigs). A gap-filling algorithm was applied to the graph to find a path from one contig to the other across the gap, consecutively using two different approaches. The first was a greedy depth-first search to find any path that fit closely to the estimated span in the scaffold, at first with a relatively tight tolerance of 100bp, and then increasing by 100bp increments until the minimum and maximum estimated spans were reached. If no path was found in this manner, the second approach was applied, which used an implementation of Dijkstra's single-source shortest-path algorithm in the networkx python package to find a path in the overlap graph that maximized the sum of alignment scores from the “source” (left of gap) contig to the “sink” (right of gap) contig[49]. If the optimal path found implied a distance between the source and sink contigs within the expected distance constraints then this path was used, otherwise the gap was filled with Ns equal to the estimated gap size. If either approach found a path, then the layout implied by the path and the reads in the path were fed as a layout message to the AMOS make-consensus algorithm to call consensus, thus filling the gap.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This study was supported in part by NIH/NIGMS grant R01GM068851 (JJM and WPR).

References

1. Chin CS, et al. The origin of the Haitian cholera outbreak strain. N Engl J Med. 2011; 364:33–42. [PubMed: 21142692]
2. Rasko DA, et al. Origins of the E. coli strain causing an outbreak of hemolytic-uremic syndrome in Germany. N Engl J Med. 2011; 365:709–717. [PubMed: 21793740]
3. Rohde H, et al. Open-source genomic analysis of Shiga-toxin-producing E. coli O104:H4. N Engl J Med. 2011; 365:718–724. [PubMed: 21793736]
4. Ali A, et al. Recent clonal origin of cholera in Haiti. Emerging infectious diseases. 2011; 17:699–701. [PubMed: 21470464]
5. Chin, C.-s., et al. The Origin of the Haitian Cholera Outbreak Strain. The New England journal of medicine. 2010:1–10.
6. Hendriksen RS, et al. Population genetics of Vibrio cholerae from Nepal in 2010: evidence on the origin of the Haitian outbreak. MBio. 2011; 2:e00157–00111. [PubMed: 21862630]
7. Mutreja A, et al. Evidence for several waves of global transmission in the seventh cholera pandemic. Nature. 2011; 477:462–465. [PubMed: 21866102]
8. Reimer AR, et al. Comparative genomics of Vibrio cholerae from Haiti, Asia, and Africa. Emerging infectious diseases. 2011; 17:2113–2121. [PubMed: 22099115]
9. Metzker ML. Sequencing technologies — the next generation. Nature Reviews Genetics. 2009; 11:31–46.
10. Schadt EE, Turner S, Kasarskis A. A Window into Third Generation Sequencing. Human molecular genetics. 2010
11. Mardis ER. Next-generation DNA sequencing methods. Annual review of genomics and human genetics. 2008; 9:387–402.
12. Chaisson M, Pevzner P, Tang H. Fragment assembly with short reads. Bioinformatics (Oxford, England). 2004; 20:2067–2074.
13. Pevzner PA, Tang H, Waterman MS. An Eulerian path approach to DNA fragment assembly. Proceedings of the National Academy of Sciences of the United States of America. 2001; 98:9748–9753. [PubMed: 11504945]
14. Myers EW. The fragment assembly string graph. Bioinformatics (Oxford, England). 2005; 21(Suppl 2):ii79–85.
15. Medvedev P, Brudno M. Maximum likelihood genome assembly. Journal of computational biology : a journal of computational molecular cell biology. 2009; 16:1101–1116. [PubMed: 19645596]
16. Batzoglou S, et al. ARACHNE: a whole-genome shotgun assembler. Genome Res. 2002; 12:177–189. [PubMed: 11779843]
17. Myers EW, et al. A whole-genome assembly of Drosophila. Science. 2000; 287:2196–2204. [PubMed: 10731133]
18. Schatz MC, Delcher AL, Salzberg SL. Assembly of large genomes using second-generation sequencing. Genome research. 2010:1165–1173. [PubMed: 20508146]
19. Chaisson MJ, Pevzner P.a. Short read fragment assembly of bacterial genomes. Genome research. 2008; 18:324–330. [PubMed: 18083777]
20. Simpson JT, et al. ABySS: a parallel assembler for short read sequence data. Genome research. 2009; 19:1117–1123. [PubMed: 19251739]

21. Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome research. 2008; 18:821–829. [PubMed: 18349386]

22. Butler J, et al. ALLPATHS: de novo assembly of whole-genome shotgun microreads. Genome research. 2008; 18:810–820. [PubMed: 18340039]

23. Margulies M, et al. Genome sequencing in microfabricated high-density picolitre reactors. Nature. 2005; 437:376–380. [PubMed: 16056220]

24. Li R, et al. De novo assembly of human genomes with massively parallel short read sequencing. Genome research. 2010; 20:265–272. [PubMed: 20019144]

25. Kingsford C, Schatz MC, Pop M. Assembly complexity of prokaryotic genomes using short reads. BMC Bioinformatics. 2010

26. Alkan C, Sajjadian S, Eichler EE. Limitations of next-generation genome sequence assembly. Nature methods. 2010

27. Chain PSG, et al. Genomics. Genome project standards in a new era of sequencing. Science (New York, N.Y.). 2009; 326:236–237.

28. Alkan C, Coe BP, Eichler EE. Genome structural variation discovery and genotyping. Nature reviews. Genetics. 2011

29. Li Y, et al. Structural variation in two human genomes mapped at single-nucleotide resolution by whole genome de novo assembly. Nat Biotechnol. 2011; 29:723–730. [PubMed: 21785424]

30. Nelson KE, et al. A Catalog of Reference Genomes from the Human Microbiome. Science. 2010; 328:994–999. [PubMed: 20489017]

31. Liolios K, et al. The Genomes On Line Database (GOLD) in 2009: status of genomic and metagenomic projects and their associated metadata. Nucleic Acids Res. 2010; 38:D346–354. [PubMed: 19914934]

32. Schadt EE, Turner S, Kasarskis A. A window into third-generation sequencing. Hum Mol Genet. 2010; 19:R227–240. [PubMed: 20858600]

33. Goldberg SMD, et al. A Sanger/pyrosequencing hybrid approach for the generation of high-quality draft assemblies of marine microbial genomes. Proceedings of the National Academy of Sciences of the United States of America. 2006; 103:11240–11245. [PubMed: 16840556]

34. Pop M. Genome assembly reborn: recent computational challenges. Briefings in bioinformatics. 2009; 10:354–366. [PubMed: 19482960]

35. Miller JR, et al. Aggressive assembly of pyrosequencing reads with mates. Bioinformatics (Oxford, England. 2008; 24:2818–2824.

36. Reinhardt, J.a., et al. De novo assembly using low-coverage short read sequence data from the rice pathogen Pseudomonas syringae pv. oryzae. Genome research. 2009; 19:294–305. [PubMed: 19015323]

37. Kong A, et al. Parental origin of sequence variants associated with complex diseases. Nature. 2009; 462:868–874. [PubMed: 20016592]

38. Ritz A, Bashir A, Raphael BJ. Structural variation analysis with strobe reads. Bioinformatics (Oxford, England). 2010; 26:1291–1298.

39. Grim CJ, et al. Genome sequence of hybrid Vibrio cholerae O1 MJ-1236, B-33, and CIRS101 and comparative genomics with V. cholerae. J Bacteriol. 2010; 192:3524–3533. [PubMed: 20348258]

40. Frerichs, RR.; Keim, PS.; Barrais, R.; Piarroux, R. Clinical microbiology and infection : the official publication of the European Society of Clinical Microbiology and Infectious Diseases. 2012. Nepalese origin of cholera epidemic in Haiti. In Press

41. Davis BM, Waldor MK. CTXphi contains a hybrid genome derived from tandemly integrated elements. Proc Natl Acad Sci U S A. 2000; 97:8572–8577. [PubMed: 10880564]

42. Rubin EJ, Lin W, Mekalanos JJ, Waldor MK. Replication and integration of a Vibrio cholerae cryptic plasmid linked to the CTX prophage. Mol Microbiol. 1998; 28:1247–1254. [PubMed: 9680213]

43. Hassan F, Kamruzzaman M, Mekalanos JJ, Faruque SM. Satellite phage TLCphi enables toxigenic conversion by CTX phage through dif site alteration. Nature. 2010; 467:982–985. [PubMed: 20944629]

44. Mazel D, Dychinco B, Webb VA, Davies J. A distinctive class of integron in the Vibrio cholerae genome. Science. 1998; 280:605–608. [PubMed: 9554855]

45. Rowe-Magnus DA, Guerout AM, Mazel D. Super-integrons. Research in microbiology. 1999; 150:641–651. [PubMed: 10673003]

46. Mazel D. Integrons: agents of bacterial evolution. Nature reviews. Microbiology. 2006; 4:608–620.

47. Pop M, Kosack DS, Salzberg SL. Hierarchical scaffolding with Bambus. Genome Res. 2004; 14:149–159. [PubMed: 14707177]

48. Kurtz S, et al. Versatile and open software for comparing large genomes. Genome Biol. 2004; 5:R12. [PubMed: 14759262]

49. Dijkstra EW. A note on two problems in connexion with graphs. Numerische mathematik. 1959; 1:269–271.

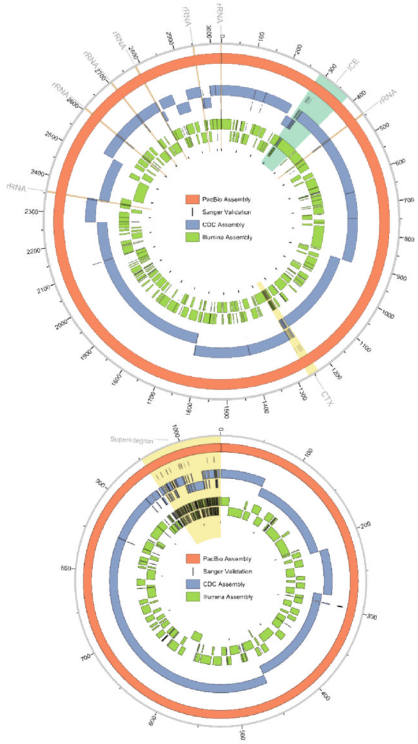

Figure 1. H1 Assembly

The completely circularized chromosomes for H1. The outermost track (salmon) represents the circularized assembly with PacBio reads. The next track indicates points of Sanger validation between CDC contigs. The middle track (blue) indicates the position of CDC contigs and the innermost track (green) shows Illumina contigs greater than 100 bp. The highlighted regions correspond to the genomic positions of the rRNA operons (Figure 2), CTX (Figure 3), superintegron (Figure 4), and ICE (Supplementary Figure 12). The origin of replication is located at 2.76 Mb in Chr1 and 632kb in Chr2.

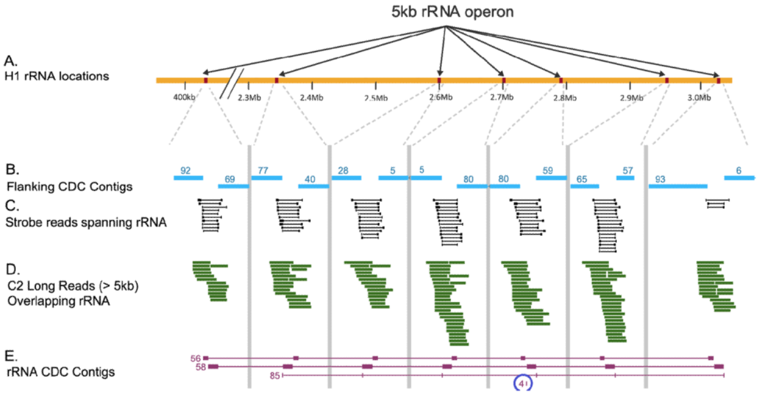

Figure 2. Resolution of rRNA genes

A) The locations of rRNA operons within H1 chromosome I. **B)** CDC contigs that flank the rRNA regions. **C)** Strobe reads that link two flanking regions, scaffolding over the 5–6kb repeat. **D)** Long reads overhanging into region allowing recalling of the rRNA repeat substructure. Here we only show the subset of reads 5kb in length that have at least 2kb of anchor sequence. **E)** CDC contigs internal to repeats. Note, that not all repeat regions contain the same constituent contigs.

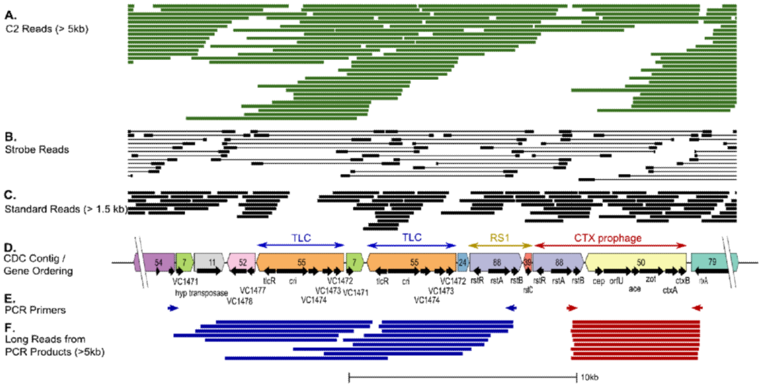

Figure 3. CTX/TLC Assembly and Validation

A) Alignment of C2 Data (> 5kb) on the CTX H1 assembly. **B)** Strobe and **C)** continuous reads were used to create an initial scaffold of the contigs within the CTX/TLC region. Concordant strobe reads (with spans between 5.5–7kb) are shown over the region. **C)** Long reads were used to fill-in gaps/resolve tandem repeat structures; selected long reads (> 1.5kb) are shown in the region. **D)** Ordering and directionality of CDC contigs (colored directed blocks) and genes (small black arrows). Each CDC contig is given a different color to highlight repeated elements. **E)** PCR primers were designed to validate the region upstream of CTX as well as the TLC structure. **F)** PCR products were sequenced and mapped back to confirm the structure; a sampling of subreads (> 5kb) that aligned to the products is shown.

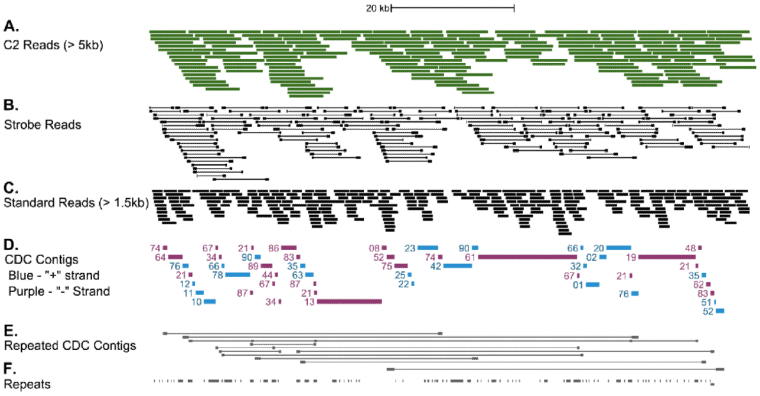

Figure 4. Superintegron Assembly

A) C2 reads **B)** strobes and **C)** continuous reads were used to scaffold and fill-in gaps across the superintegron region. The complexity of the region is highlighted by the number of CDC contigs in the region (**D,E)**. **D)** The contigs scaffolded together – blue indicates positive strand mappings, purple indicates negative strand. **E)** Contigs that are repeated – with linkages between the repeated positions. **F)** Shows repeats as identified by nucmer. Note, not all repeated contigs necessarily are repeats as contigs may only be present in truncated forms.

Table 1

Sequencing Statistics for H1[5]

Dataset	Number of reads	Mapped coverage	Mean read length	Mean read accuracy
Illumina 1×36bp	28.6 M	244X	36 bp	99.9
454	248 K	20X	329 bp	96.0
PacBio Standard Continuous Reads	68565	19X	1.28 kb	85.1
PacBio Strobe Reads	17106	26X*	6.180 kb	84.7
PacBio C2 Reads	248171	138X	2.5kb	84.9

*Mapped *physical* coverage for strobe reads including the strobe span length.

Table 2

Assembly Statistics for H1

Dataset	Number of scaffolds >1kb	Total number of scaffolds	Number of scaffolds covering 99% of genome	N50	Total number of contig	Total number of s N's in scaffolds	Consensus Accuracy
Illumina *	313	2594	416	20 kb	2594	0	99.99
lllumina + 454 *	65	93	45	155 kb	93	0	99.99
Illumina + 454 + PacBio continuous reads	32	45	20	638 kb	49	1229	99.99
Illumina + PacBio continuous reads (Standard long + strobe)	12	2110	6	1.10 Mb	2617	105kb	99.99
Illumina+ 454 + PacBio continuous reads (Standard long + C2) + PacBio strobe *	2	2	2	3.01 Mb	2	0	99.99

*The contigs indicated for these assemblies contain no gaps (no uncalled bases). The other assemblies may contain uncalled bases, but such bases are supported by at least one read, but with not enough coverage to realize high accuracy consensus base calls.