

Enhanced De Novo Assembly of High Throughput Pyrosequencing Data Using Whole Genome Mapping

Fatma Onmus-Leone^{1*}, Jun Hang^{2*}, Robert J. Clifford¹, Yu Yang², Matthew C. Riley¹, Robert A. Kushner², Paige E. Waterman¹, Emil P. Lesho¹

1 Multidrug-Resistant Organism Surveillance Network and Repository, Walter Reed Army Institute of Research, Silver Spring, Maryland, United States of America, **2** Viral Diseases Branch, Walter Reed Army Institute of Research, Silver Spring, Maryland, United States of America

Abstract

Despite major advances in next-generation sequencing, assembly of sequencing data, especially data from novel microorganisms or re-emerging pathogens, remains constrained by the lack of suitable reference sequences. *De novo* assembly is the best approach to achieve an accurate finished sequence, but multiple sequencing platforms or paired-end libraries are often required to achieve full genome coverage. In this study, we demonstrated a method to assemble complete bacterial genome sequences by integrating shotgun Roche 454 pyrosequencing with optical whole genome mapping (WGM). The whole genome restriction map (WGRM) was used as the reference to scaffold *de novo* assembled sequence contigs through a stepwise process. Large *de novo* contigs were placed in the correct order and orientation through alignment to the WGRM. *De novo* contigs that were not aligned to WGRM were merged into scaffolds using contig branching structure information. These extended scaffolds were then aligned to the WGRM to identify the overlaps to be eliminated and the gaps and mismatches to be resolved with unused contigs. The process was repeated until a sequence with full coverage and alignment with the whole genome map was achieved. Using this method we were able to achieve 100% WGRM coverage without a paired-end library. We assembled complete sequences for three distinct genetic components of a clinical isolate of *Providencia stuartii*: a bacterial chromosome, a novel *bla*_{NDM-1} plasmid, and a novel bacteriophage, without separately purifying them to homogeneity.

Citation: Onmus-Leone F, Hang J, Clifford RJ, Yang Y, Riley MC, et al. (2013) Enhanced De Novo Assembly of High Throughput Pyrosequencing Data Using Whole Genome Mapping. PLoS ONE 8(4): e61762. doi:10.1371/journal.pone.0061762

Editor: Jingfa Xiao, Beijing Institute of Genomics, China

Received: September 28, 2012; **Accepted:** March 11, 2013; **Published:** April 17, 2013

This is an open-access article, free of all copyright, and may be freely reproduced, distributed, transmitted, modified, built upon, or otherwise used by anyone for any lawful purpose. The work is made available under the Creative Commons CC0 public domain dedication.

Funding: This study was supported by the United States Army Medical Command Policies 09-050 and 11-035, and was partially funded by grants C0709_12_WR and I0361_12_WR from the Global Emerging Infections Surveillance and Response System, a Division of the Armed Forces Health Surveillance Center. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: Fatma.Onmus-Leone.ctr@us.army.mil (FOL); Jun.Hang.ctr@us.army.mil (JH)

These authors contributed equally to this work.

Introduction

Next-generation sequencing (NGS) technologies, which enable the rapid generation of whole genome sequences, have revolutionized genomic research [1–3]. With the emergence of low-cost bench top NGS platforms, genomic studies are now being performed in translational or applied research labs instead of state-of-the-art genome centers [4,5]. Despite these technical advances, complete *de novo* sequence assembly and finishing (genome closure) continue to challenge scientists [6]. Consequently, a large number of unfinished genome draft sequences have been submitted to public databases [7]. Streamlined approaches for assembling high-quality full genome sequences are lacking, especially in the multidrug-resistant Gram-negative microbial genomics field, where *de novo* sequencing is frequently required due to the genetic diversity and dynamic genome rearrangements that occur in these organisms [8–10].

Among the various NGS platforms, Roche 454 GS systems are widely used for sequencing bacterial genomes [11]. The long read lengths - 300–500 bp for the FLX system and 500–800 bp for the new FLX+ system - and high sequence quality enable *de novo* assembly of reads from shotgun sequencing into large contigs.

These contigs can be used for gene annotation, genotyping, and phylogeny. However, when a complete genome sequence is desired, distance information, such as that from paired-end library sequencing or extra-long read length sequencing using PacBio RS system (Pacific Biosciences, Menlo Park, CA) is required to link the *de novo* contigs [12].

WGM uses single-molecule restriction analysis to obtain information about the sizes of the restriction fragments and their physical positions along the DNA strand [13]. WGM has been used in a variety of applications, which include genotyping and phylogenetic analyses of related microbial isolates [14–23], detection of large genomic structural variations or rearrangements [24–26], and verification or quality control for assembled genome sequences [27–34]. A restriction-based physical genome map also has the potential to serve as a reference for the accurate ordering of NGS contigs and to facilitate closing the gaps between mapped contigs.

To date, the strategy of using WGM to scaffold sequencing contigs during *de novo* assembly has been used in projects that either employed multiple NGS platforms or used both paired-end and shotgun sequencing libraries on the Roche 454 platform [4,26,32,34,35]. Here we demonstrate that WGM can be used to

direct the scaffolding of *de novo* contigs from the Roche 454 GS shotgun library sequencing to achieve the complete genome sequence of a clinical isolate of *Providencia stuartii* for which no reference genome sequence was available. This work produced the first complete *Providencia stuartii* genome sequence deposited in GenBank. WGM directed NGS assembly also facilitated the detection of extrachromosomal structures in *Providencia stuartii* MRSN 2154; namely a plasmid 178 kb in length, and a novel bacteriophage. It also yielded complete nucleotide sequences for both [36,37].

Materials and Methods

Clinical bacterial isolate

MRSN 2154 is a multidrug-resistant isolate of *Providencia stuartii* recovered from the blood of a patient in Afghanistan, who ultimately succumbed to multiple infectious complications. It was subsequently found to harbor a large plasmid carrying multiple antibiotic resistance genes, including *bla*_{NDM-1}, the New Delhi metallo- β -lactamase gene 1 [37,38].

Whole genome mapping (WGM)

WGM using the Argus system (OpGen Inc, Gaithersburg, MD) involves four steps: DNA extraction; immobilization and *in situ* restriction digestion; image capture and measurement; and map assembly and analysis. MRSN 2154 was grown on blood agar at 37°C overnight. DNA was isolated from a single colony approximately 2 mm in diameter using the Argus Sample Preparation Kit in conjunction with the Agencourt Genfind v2 DNA Isolation Kit as described by the manufacturer (OpGen Inc). This whole cell lysis protocol allows chromosomal and extrachromosomal structures to be coextracted in the same sample preparation. DNA quality and quantity was assessed using the Argus QCard Kit. In the quality control step, DNA molecules were stretched along the QCard surface, stained using Argus Stain Kit, and visualized using Argus imaging system. After verifying DNA concentration (approximately 10 molecules per frame) and length (greater than 200 kb per molecule), the sample was mapped using the Argus MapCard Kit. Briefly, single DNA molecules migrating through microfluidic channels by capillary action were immobilized on a charged glass surface. Using a restriction enzyme (NcoI) predetermined by the Argus Enzyme Chooser software module, immobilized DNA was digested *in situ* in a MapCard Processor. The Argus Enzyme Chooser software identifies optimal restriction enzymes used for WGM by performing a series of virtual restriction digests of a reference genome sequence for the bacterial species or, in the case of a species without a reference genome sequence, *de novo* sequencing contigs. The software chooses an enzyme that produces an average restriction fragment size of 6–12 kb and a maximum fragment size less than 80 kb. After digestion, molecules were imaged using fluorescence microscopy on the Argus WGM system and automatically processed using its image acquisition software (OpGen Inc). This software detects cut sites in DNA molecules and calculates the length of the resulting restriction fragments by using a pixel length to kilobase conversion algorithm.

Iterative assembly from single molecules to the consensus WGRM is performed by the Argus assembly software (OpGen Inc). Downstream analysis (sequence placement) was carried out using Argus MapSolver software. DNA sequence was imported into MapSolver™ software and converted into *in-silico* maps using the same restriction enzyme as was used to generate the respective WGRM. DNA sequence contigs were aligned to the WGRM using the sequence placement function of MapSolver, which uses a

dynamic programming algorithm that finds the optimal alignment of two restriction maps according to a scoring model that incorporates fragment sizing errors, false and missing cuts, and missing small fragments. The algorithm applies user-provided settings toward generating local alignments between each contig and the WGRM. DNA sequence contigs are aligned in both forward and reverse directions. MapSolver generates an alignment score for each comparison, where a higher score implies greater confidence in the alignment [39]. Alignment scores are calculated by awarding aligned fragments from the WGRM and the *in silico* map that have the exact same size in base pairs with a value of 1 and are penalized when the size or pattern deviates. Restriction fragment patterns that align will result in scores that are additive, based on the aligned fragments they contain. Therefore, longer alignments between more similar restriction patterns produced higher scores. It is recommended to use the default score of 3 for the initial replacement process. The score can be increased up to 6 to filter out spurious alignments or decreased to identify all potential alignments.

High throughput pyrosequencing and *de novo* sequence assembly

Extraction of total DNA and high throughput pyrosequencing using the Roche GS FLX Titanium system (Roche 454 Life Sciences, Branford, CT) were previously described in detail [37]. In brief, 5 μ g total DNA purified using PurElute Bacterial Genomic Kit (EdgeBio, Gaithersburg, MD) was fragmented by

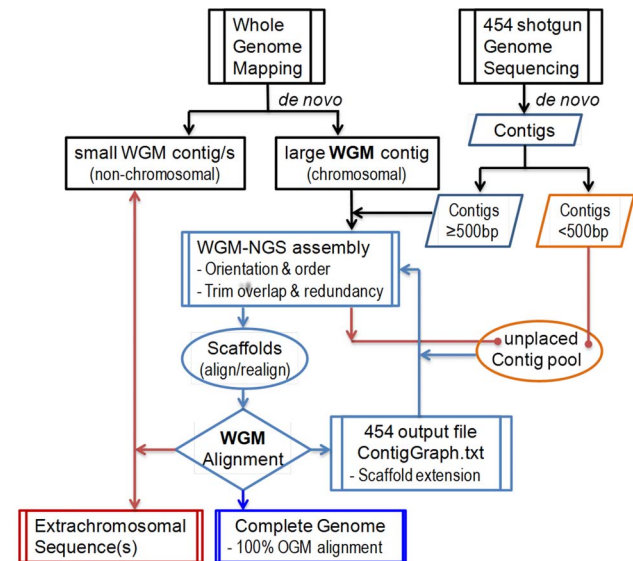


Figure 1. Overview of the WGM-NGS *de novo* sequence assembly process. WGM and Roche 454 pyrosequencing data are *de novo* assembled respectively. WGM assembly software aligns single molecule restriction maps of chromosomal DNA fragments to form a contiguous contig. Depending on the size and number of the cut sites WGM can also assemble contigs for non-chromosomal elements. Roche 454 contigs that align (*in-silico* mapping) to the large WGRM contig (physical reference) are further assembled into scaffolds. Gaps are filled by extending scaffolds with sequences in the unplaced contig pool based on contig branching structure information (454 contig graph). Extended scaffolds are subsequently realigned to the WGRM. The process is repeated until high restriction pattern similarity and WGRM coverage is achieved. Contigs that do not assemble the genome are compared with the WGRM chromosomal assembly to confirm the existence of the extrachromosomal content. doi:10.1371/journal.pone.0061762.g001

adaptive focused acoustic forces using the Covaris S2 system, with the shearing protocol for producing DNA fragments around 1 kb in length (Covaris, Woburn, MA). The shotgun rapid ligation library of DNA fragments was prepared, and then subjected to size selection using a 2% E-gel (Invitrogen, Carlsbad, CA) to isolate the fraction of DNA with sizes ranging from 600 to 800 bp. The library was sequenced, and the reads were subjected to *de novo* assembly using the GS Assembler software (Newbler) version 2.5.3. All contigs were used in the final sequence assembly.

WGM-oriented assembly of scaffolds and finishing of genome assembling

Whole genome restriction map assembly software aligns single molecule restriction maps of chromosomal DNA fragments, to form a contiguous contig (large contig). Molecules that are not aligned to the chromosomal map assemble as smaller contigs with smaller mapsets (usually in upper teens in number). These contigs can contain chromosomal molecules which do not match the genomic map restriction pattern, in addition to extra-chromosomal elements such as plasmids. A whole-genome NcoI restriction map was used as a template to arrange *de novo* contigs in the correct order and orientation. The workflow illustrated in Figure 1 contains the following steps: (1) The MapSolver program was used to generate *in silico* NcoI restriction maps of all NGS *de novo* contigs and align them with the MRSN 2154 WGRM. As a result, contig alignments, gaps, overlaps, and redundant contigs on the physical map were identified. (2) To fill the gaps, aligned contigs were extended using contig branching structure information from the *de novo* assembly output file 454ContigGraph.txt. The contig graph file describes the relative orientation of adjacent contigs. As an example, contig 13 has contig 43 at its 3'-end (C13→43), contig 43 has contig 12 or 31 at its 5'-end, and contig 13 or 42 at its 3'-end (12/31→C43→13/42). The contig graph

information also supports the contig relationships 32→C31→43 and 18/26→C32→8/31. Therefore, the sequence between two WGM-placed contigs 13 and 18 was determined to be 13→43RC→31RC→32RC→18, where RC stands for reversed and complementary (Figure 2B–C). Additionally, contigs which did not align to the WGRM due to an inadequate number of cut sites were tested to fill the gaps. After gaps in the scaffold were filled, GS Assembler reference mapping was used to confirm continuous genome coverage across the resulting merged contigs. (3) The assembled scaffolds were aligned with the WGRM to ensure the assembly, and then joined together by eliminating any overlapped terminal sequences. (4) The assembled draft genome sequence was aligned with the WGRM to identify discordant regions. Unmapped contigs were fit into these discordant regions, and their placement was confirmed by re-alignment to the WGRM, after merging them together. Steps 1 through 4 were repeated until full genome coverage was achieved. Two additional steps (not shown in Figure 1) were taken to finish the genome assembly: Firstly, PCR amplification and Sanger sequencing were used to resolve the structure and organization of the multi-copy ribosomal RNA operons. Secondly whole-genome gene-coding annotations were used to correct reading frame shifts caused by sequencing errors in homopolymer regions. Geneious Pro v4.7 (Biomatters Ltd, Auckland, New Zealand) [40] was used for sequence editing, concatenation, translation, *in silico* restriction, and sequence alignment.

Results

WGM technology has been used to generate high resolution whole-genome restriction map for global genomic comparison among closely related strains of microbes [41,42]. In this study, we integrated WGM with NGS shotgun sequencing to streamline

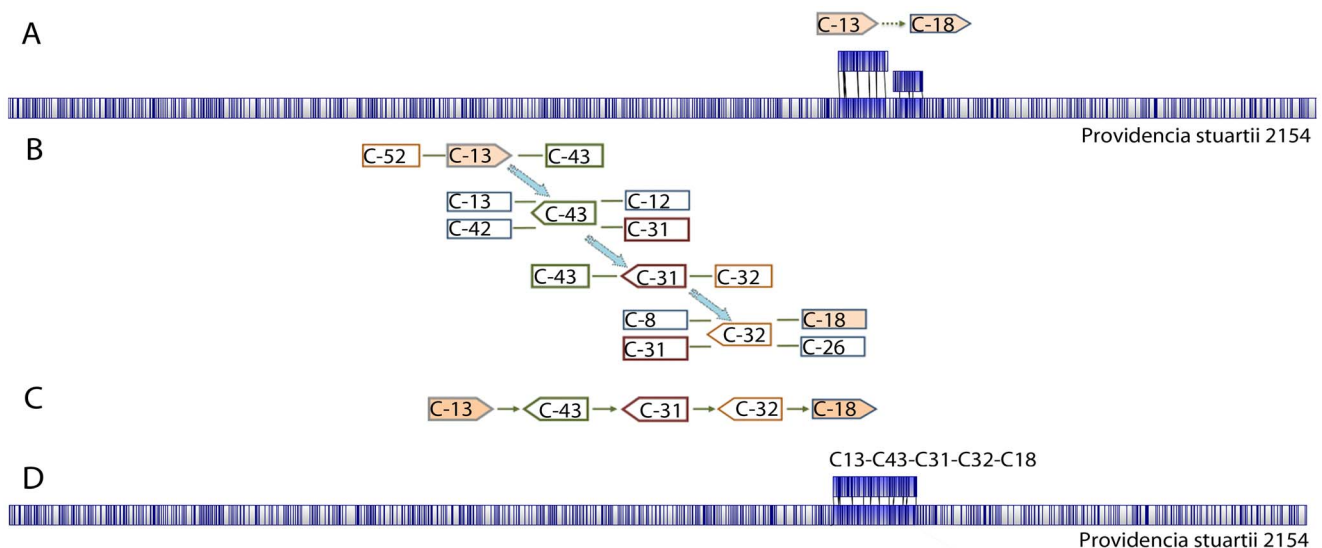


Figure 2. Integration of WGM alignment and 454 contig graph information for sequence assembly gap filling. The MapSolver alignment software was used with default parameters for *in-silico* alignment of sequence contigs to the WGRM. Unaligned regions are highlighted in white while alignment regions are shaded blue. Vertical lines represent NcoI cut sites. (A) Contigs 13 and 18 were *in-silico* digested and aligned to the WGRM to determine their orientations and the distance that separates them. There was a gap between contigs 13 and 18. (B) Contig branching information was used to fill the gap with contigs 43, 31, and 32. Contigs represented by boxes with right-pointing arrowheads are in the forward orientation relative to the WGRM; those represented by boxes with left-pointing arrowheads are in the reverse orientation; those represented by rectangular boxes are of unknown orientation relative to the WGRM. (C) Concatenation of contigs which were ordered and oriented by the WGM-enhanced gap-closure process. (D) Mapping the concatenated sequence to the WGRM to verify the quality of the gap filling. An *in-silico* digest of a contig is shown above the WGRM.

doi:10.1371/journal.pone.0061762.g002

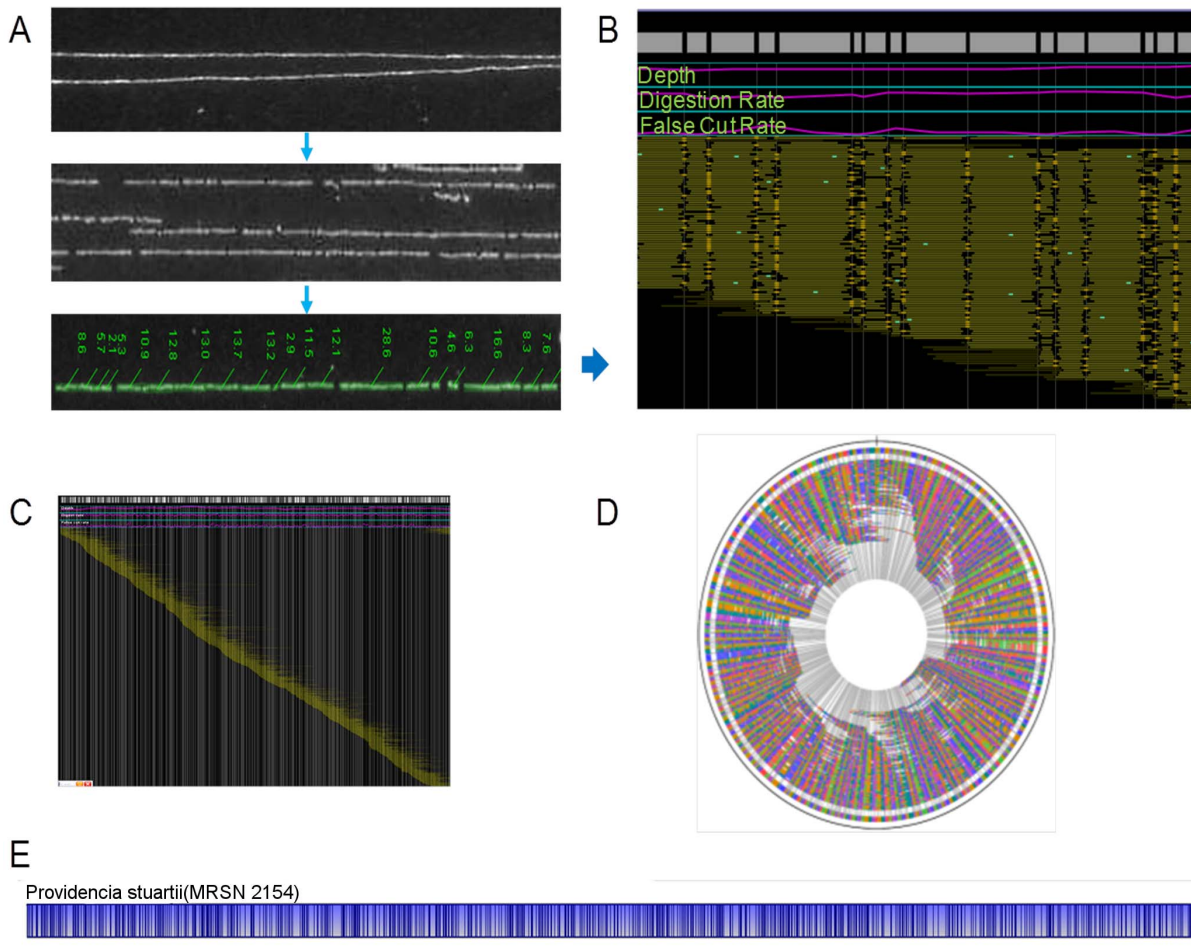


Figure 3. Construction of a whole-genome restriction map for *Providencia stuartii* MRSN 2154. (A) Genomic DNA was immobilized, *in situ* digested with NcoI enzyme, measured and converted to a digital profile. Black gaps within single DNA molecules are breaks created by restriction digestion. (B) DNA molecule restriction patterns were aligned to generate WGRM contigs. Each green horizontal line represents a single DNA molecule. Gray blocks represent the consensus DNA restriction fragments. Black vertical lines represent consensus restriction cut sites. (C) WGRM contigs aligned to create continuous coverage along the DNA strand, producing the whole-genome consensus map. (D) Circular representation of the linear consensus map which has an estimated size of 4.2 Mb. (E) The MRSN 2154 consensus genome map displayed by the MapSolver program. The circular map is illustrated in a linear view. NcoI restriction sites are indicated as vertical bars.
doi:10.1371/journal.pone.0061762.g003

complete genome assembly of a clinically relevant bacterial pathogen.

Whole Genome Mapping of *P. stuartii* MRSN 2154

After the initial phenotypic and molecular investigation of MRSN 2154, which was the first reported *bla*_{NDM-1} containing *P. stuartii* strain [37], we applied WGM to characterize the isolate on a genomic scale. The WGM images revealed the presence of three distinct types of DNA molecules: high molecular weight chromosomal DNA fragments (Figure 3), circular plasmid molecules, and linear DNA molecules with a repetitive restriction pattern (Figure 4). A circular WGRM chromosomal assembly, with coverage ranging from 52 to 117 fold, was assembled from restriction maps of 1179 DNA molecules. The estimated genome size is 4.2 Mb (Figure 3). The plasmid was not assembled in WGM. It was estimated to be approximately 160 kb, which is similar to the sizes of some NDM-1 bearing plasmids seen in other Gram-negative bacteria [43–45]. The repetitive DNA structure was not assembled into the chromosome map. The coverage of

this distinct repetitive element ranged from 30 to 50 fold. The size of the repeat unit was approximately 54 kb (Figure 4).

Roche 454 sequencing and assembly of complete sequences

A shotgun rapid ligation library for MRSN 2154 was sequenced using the Roche GS FLX Titanium system. 274,932 filtered reads, totaling 82 Mb of sequence underwent *de novo* assembly using the GS Assembler software (Newbler) version 2.5.3. In total, 70 contigs ≥ 100 bp in size and 11 smaller contigs were assembled, with the average coverage depth being 17.5-fold (Table S1). Among the 81 *de novo* assembly contigs, 19 were assembled into the 178 kb plasmid pMR0211, which carries the *bla*_{NDM-1} gene (JN687470, NC_016974). Assembly and annotation of the pMR0211 sequence were described elsewhere [37].

The WGRM of the MRSN 2154 chromosome was used to guide assembly of contigs into the complete genome sequence (see Methods; Figure 5). Contigs 1–18 and 20, which provided 89% genome map coverage, were ordered and oriented along the WGRM (Figure 5A). The process provided information on

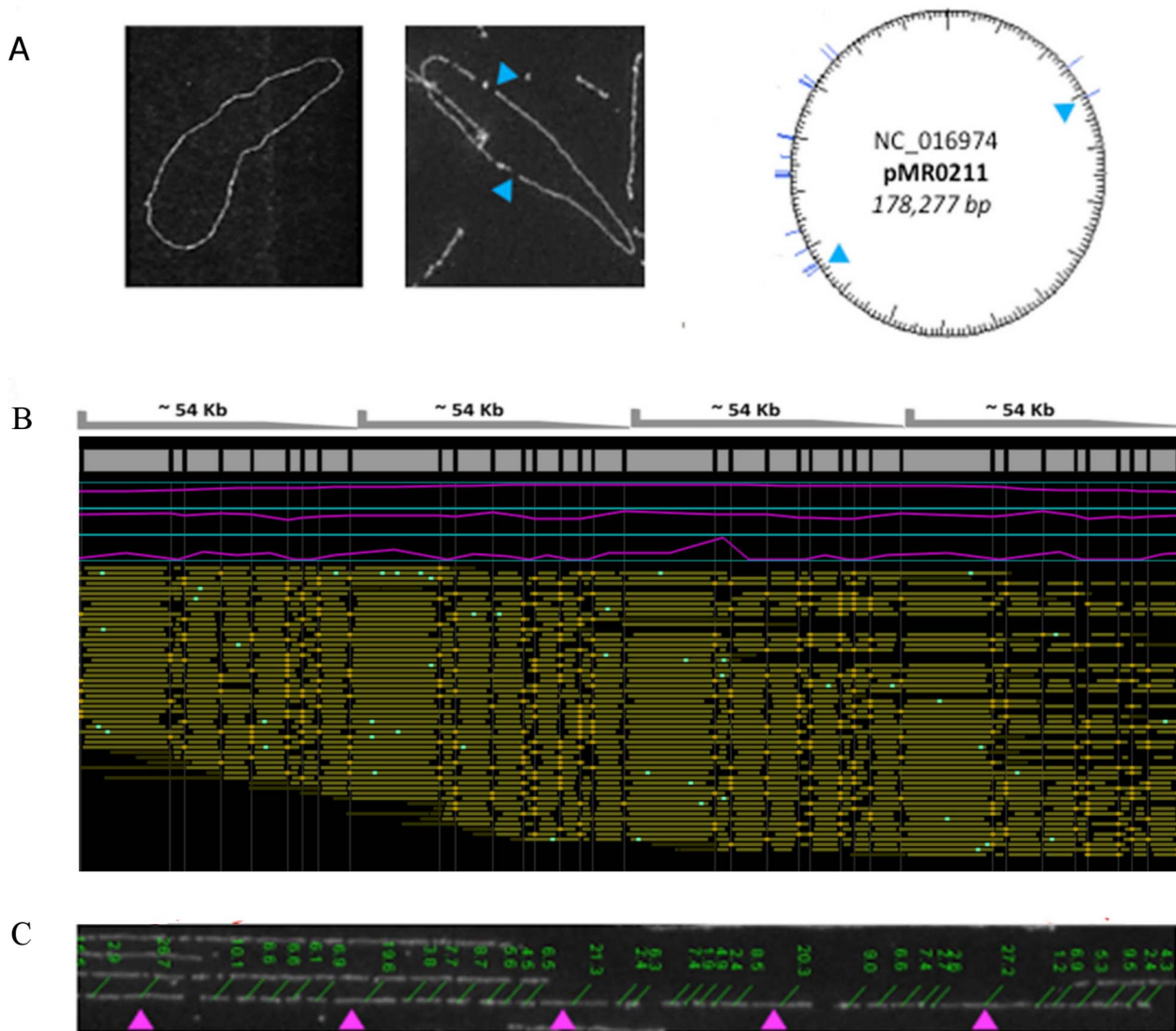


Figure 4. Whole genome mapping of two extra-chromosomal DNA elements in *Providencia stuartii* MRSN 2154. (A) A large circular plasmid with its estimated size (>160 kb) and NcoI pattern consistent with the *bla*_{NDM-1} plasmid carried by MRSN 2154. *Left*, the plasmid prior to NcoI digestion (QCard image); *center*, the NcoI-digested plasmid (MapCard image); *right*, the *in silico* NcoI restriction map for the *bla*_{NDM-1} plasmid. (B) A novel putative bacteriophage coexisting with MRSN 2154. The Consensus WGRM assembled for the bacteriophage suggests a repetitive structure; each repeat is approximately 54 kb in size. (C) Sized image of one of the DNA molecules from the bacteriophage assembly. The 20 kb NcoI fragments are indicated with the arrows in the image.
doi:10.1371/journal.pone.0061762.g004

redundant contigs, overlaps between adjacent contigs, and the approximate physical distances of gaps that required filling or closure with unplaced contigs. Redundant contigs and the overlapped terminal sequences were then removed. For the unfilled regions, the two contigs upstream and downstream of each gap were extended using contig branching structure information in the 454 *de novo* assembly contig graph file. After gap closure, the resulting 8 scaffolds were aligned to the WGRM (Figure 5B). By removing overlapping sequences, we joined these scaffolds together into a draft sequence covering 99% of the genome. Using default parameters, the draft sequence was then aligned with the WGRM for manual visual alignment. A discordance was found in one region where about 32 kb of the draft sequence was missing relative to the WGRM (Figure 5D). This disagreement between 2,437,007 bp and 2,438,048 bp was corrected using two relatively large contigs which had not been

incorporated into the assembly, [contig 28 (25,715 bp) and contig 38 (4431 bp)], and the associated smaller contigs. The resulting corrected sequence showed a high level of similarity with the optical genome map. (Figure 5D).

Determination of ribosomal RNA operons

Many bacterial genomes carry multiple *rm* operons, which each contain one copy of the 5S, 16S, and 23S ribosomal RNA genes and a unique intergenic spacer region (ITS) [46]. We observed 7 repetitive structures in MRSN 2154 genome, which all contained the 3 rRNA genes and variable intermediate sequences (Figure 6A). However, neither WGM nor 454 sequencing can resolve the structures. The sizes of the *rm* operons were too long to be read through by either 454 or Sanger sequencing, and the sequences were not large enough and/or the NcoI restriction patterns were not unique enough to be distinguished by WGM. We designed

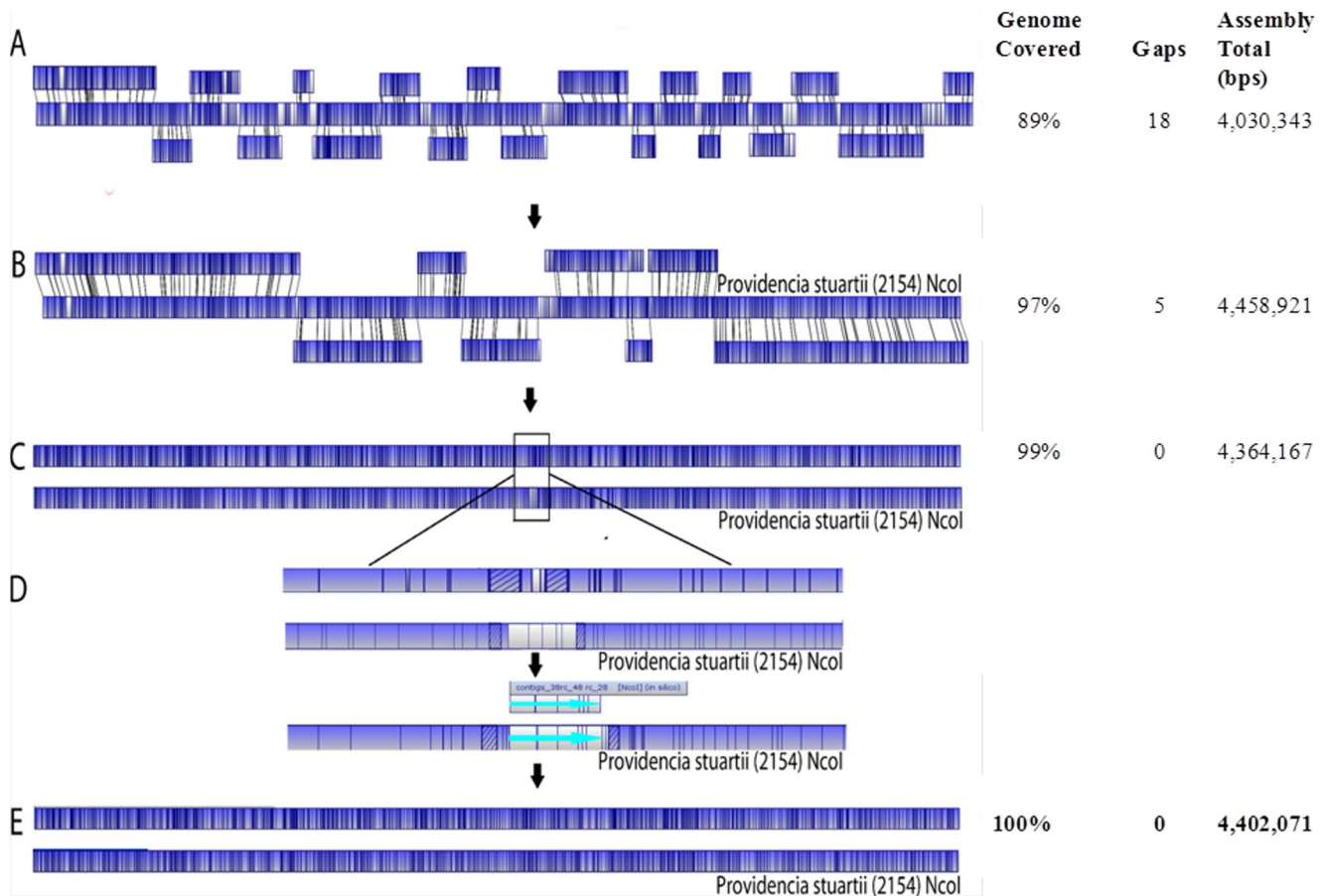


Figure 5. *Providencia stuartii* MRSN 2154 genome sequence assembly using the WGM-NGS approach. The MapSolver program highlights unaligned regions in white while aligned regions are shaded blue. (A) Contig replacement: 19 Roche 454 *de novo* contigs (>500 bp) were aligned to the MRSN 2154 WGRM in the correct orientation and order. Overlaps, redundancies, orientation and unaligned contigs were confirmed and resolved where needed. (B) Scaffolding by merging and extension: with the guidance of WGM and the branching structure information (454 contig graph), contigs were progressively joined to form scaffolds. Scaffolds were aligned to the WGRM to examine orientation, order, overlaps and gaps. (C) Alignment of first draft to the WGRM: after removal of redundant overlapping sequences, scaffolds were joined to form the first genome draft. (D) Correction of mismatches: An unaligned region of approximately 30 kb was identified (shown as white region) and subsequently resolved with the aid of the WGRM and branching structure information is shown in detail in Figure 2. Identified contigs were integrated in the draft. (E) Alignment of the complete genome sequence to the WGRM. The WGRM reference was 100% covered with highly similar NcoI restriction pattern. doi:10.1371/journal.pone.0061762.g005

primers targeting upstream and downstream sequences flanking the *mm* operons to amplify the entire *mm* region (approximate 5.5 kb), and then we sequenced in both the forward and reverse directions from the internal rRNA gene (Table 1, Figure 6B). Among the 7 *mm* operons in MRSN 2154, there are 4 different ITS sequence structures: an ITR in *mm* 1; an identical ITS in *mm* 2, 5, 6; an ITS in *mm* 3; and an identical ITS in *mm* 4 and 7 (Figure 6C).

Finally, the complete *P. stuartii* MRSN 2154 genome sequence was annotated using the NCBI Prokaryotic Genomes Automatic Annotation Pipeline (<http://www.ncbi.nlm.nih.gov/genomes/static/Pipeline.html>) and manually corrected to eliminate open reading frame shifts caused by ambiguous homopolymer calling (data not shown). After the above assembly process, a single contig, contig 22, had not been assembled into either the *P. stuartii* genome or plasmid pMR0211. To investigate whether this 58,104 bp contig is a genetic element distinct from the assembled chromosomal and plasmid genomes, the sequence was *in silico* digested with NcoI and compared with the extra-chromosomal repetitive element found by WGM. The NcoI digest of the sequence closely matched the repeat structure (Figure 4B). In addition, the contig branching information showed that this

sequence did not connect to any other contigs. Furthermore, the 454Contigs.ace file, which illustrates the alignment of the reads, revealed that the sequence is continuous end-to-end. Therefore, it is either a circular DNA or a continuous sequence of tandem repeats. The sequence was then subjected to BLASTN alignment with the GenBank non-redundant nucleotide database and to BLASTX alignment with the GenBank non-redundant protein sequences database. Both results suggested the sequence represents a novel bacteriophage, which has a nucleotide sequence similarity of 65–70% with the *Enterobacter* phage Enc34 (JQ340774.1) over 61% of its length. The evidence thus suggested that there was a bacteriophage co-existing with the clinical isolate MRSN 2154. Annotation of the bacteriophage genome was performed using the xBASE server [47] and BLASTP.

Deposition of nucleotide sequences in GenBank

The assembled complete sequences for *Providencia stuartii* MRSN 2154 genome, the *bla*_{NDM-1} plasmid, and the putative bacteriophage were deposited in GenBank under accession numbers

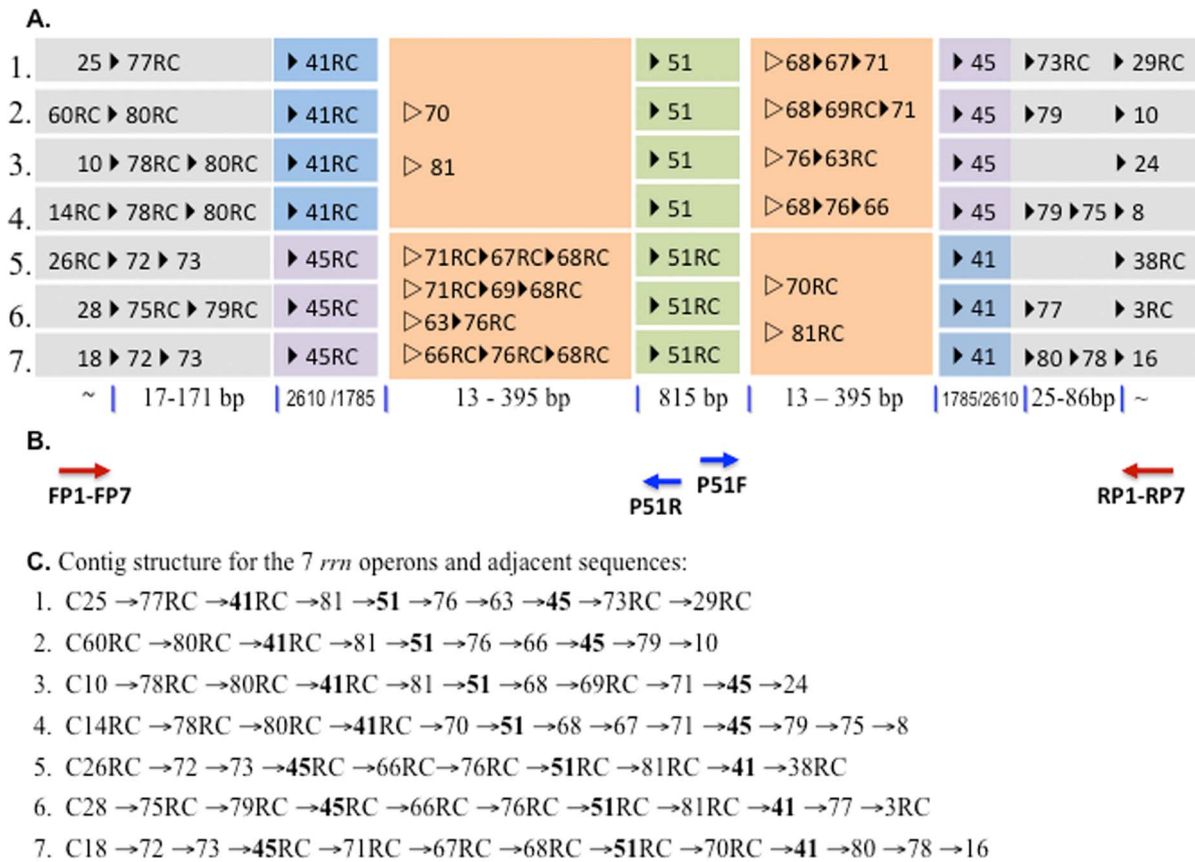


Figure 6. Resolved sequence structures for the 7 *rrn* operons in MRSN 2154. PCR and Sanger sequencing were used to determine the variable regions in *rrn* operons. RC, reversed and complementary. (A) Illustration of contig assembly for the 7 *rrn* operons and adjacent sequences. Variable regions between rRNA sequences (contigs 41, 45 and 51) were shown in orange boxes and indicated as orientation-undecided by the empty arrow. The sizes for the ribosomal and intergene regions are shown. (B) Schematic representation of regions to be amplified and the Sanger sequencing directions. Seven *rrn* regions were amplified by PCR using 7 pairs of primers, FP1-FP7 and RP1-RP7 which are corresponding to specific sequences flanking the *rrn* regions. Sanger sequencing outward from the conserved rRNA sequence (contig 51) in both directions with primers P51R and P51F was used to determine the intergene spacer regions. (C) The results for *rrn* operon sequence structures resolved by PCR amplification and Sanger sequencing.

doi:10.1371/journal.pone.0061762.g006

CP003488 and NC_01773, JN687470 and NC_016974, and JX296113 respectively.

Discussion

In this study, WGM provided both global and fine scale genomic structures that facilitated *de novo* assembly of complete bacterial genome sequences using rapid shotgun Roche 454 sequencing. We identified and sequenced three full genomic structures from a clinically important bacterial isolate. Notably, the two extrachromosomal elements, a 178-kb plasmid and a 58-kb bacteriophage, were sequenced and fully assembled without being purified to homogeneity. The quality of the assembly and the correct assignment of contigs in each genome were confirmed not only by the alignment with the WGRM profiles (Figure 5) but also by the consistency of sequence coverage depth for each contig versus the actual copy number in the final sequences (Table S1).

We demonstrated that the use of WGM simplified and streamlined NGS data assembly, by sorting and orientating *de novo* assembly contigs in the proper physical order. For laboratories that own the Argus system and can run multiple samples on a single map card [48], our technical approach is useful and cost-effective. The method eliminates the need for time-consuming and

sophisticated paired-end sequencing, and it is functional with limited bioinformatics resources. WGM laboratory processes, image acquisition, data processing, and map assembly can be done in one to three days. In contrast, it may take several days to construct one or more paired-end libraries which can link sequences with a 3 kb, 8 kb or 20 kb span. WGM directed sequence assembly provides multiple enhancements over traditional methods: (1) it facilitates the identification of extrachromosomal DNA elements; (2) it provides an assessment of DNA molecule size and configuration (circular or linear, with a repetitive pattern for repeats); (3) it enables the detection of structural variations (large-scale rearrangements, inversions, tandem repeats) relative to reference sequences; (4) it provides clear information regarding the position and size of gaps and mismatches in the assembly; and (5) it produces a whole-genome restriction map to serve as a reference for verification and quality control of the genome assembly. WGM-NGS represents a feasible technical approach for interrogating microbial genomes from delineation of genomic components and a landscape for construction of a scaffold and finishing the complete sequence.

Next-generation sequencing platforms will become increasingly available to all types of investigators and laboratories, including clinical, non-research laboratories, as the price and the operating

Table 1. Oligos for amplification of *rrn* regions 1–7 and resolving *rrn* operon structures by Sanger sequencing.

<i>rrn</i> operon	Oligo	Sequence	nucleotide position	
1	C14RC-3'F	AACCCGATGATCACAAAACCGTGT	2181743	2181766
	C8-5'R	AGCGCGAACGAAAAGAAAATGCAAA	2187870	2187846
2	C26RC-3'F	AGCCGCCGAAGGGTTCAATAGG	2433678	2433699
	C38RC-5'R	TTCGAACTGGGGACCTCACCT	2439490	2439469
3	C28-3'F	CGACAAACCGCTTCGGTGGTCA	2187795	2187774
	C3RC-5'R	AAGTGGCGGAACGGACGGGA	1192630	1192649
4	C25-3'F	AAGTGGCGGAACGGACGGGA	1192630	1192649
	C29RC-5'R	TCTGACTCCCTATAATGCGCTCC	2433783	2433806
5	C60RC-3'F	GGCGTATGTTCCGCAACGGC	1621186	1621205
	C10-5'R	AACCGTTCGGTGGTCAGGT	2187790	2187771
6	C18-3'F	ACATGCAAGCAGGAAAGCAGCA	3230145	3230167
	C16-5'R	TGCAGTTCGCTGGCACAGGAT	3236226	3236206
7	C10-3'F	GGGAAAAGAAGGGGAGTGATGGAAGGT	1809355	1809381
	C24-5'R	AACCGCTTCGGTGGTCAGGT	2187790	2187771
Sanger sequencing	C51R	GGGGACCATCTCCAAGGCT	2184810	2184790
	C51F	CACCTGCTCGCTCATGCCA	3232696	3232677

Each oligo name indicates a contig, contig orientation, terminus and extension direction. RC, reversed and complementary. Sequencing primers were designed to extend outwards, i.e. to sequence the regions upstream and downstream of contig 51.

doi:10.1371/journal.pone.0061762.t001

costs continue to decrease [5]. The procedure we described used the Roche GS FLX Titanium system and simple shotgun sequencing. With the long read length from this pyrosequencing platform, fewer than 100 *de novo* assembly contigs can be expected for a bacterial genome. With the new FLX+ system, the read length will almost double, so we expect a smaller number of contigs and a shortened time for the genome finishing process. However, to take full advantage of this technology, strategies are needed to expedite or enhance final assembly and quality in order to facilitate sensitive and precise detection of genomic variations [4].

Interestingly, the total DNA extract for the novel clinical isolate MRSN 2154 contained three distinct genetic structures at approximate equal copy number (Table S1). With the combination of WGM and NGS, we were able to sequence them altogether in a single run and assemble all three complete sequences. In our study, fine adjustments during contig scaffolding were accomplished through the combined use of OpGen MapSolver v3.2.3, Newbler assembler, and Genious and NCBI BLAST tools. A limitation of this approach is that labor-intensive and error-prone manual manipulations are still required. Further improvement will necessitate an integrated bioinformatics approach to streamline and standardize the WGM-directed sequence assembly [4,35]. Such an approach could consist of first round placement of unique contigs using the WGRM as reference, second round placement of contigs with lower alignment scores to fill the unaligned WGRM regions, then gap closure based on contig branching structure, followed by merging overlapped sequences, and lastly, fine adjustment at joint regions by sequence remapping using original reads. Using additional enzymes in WGM to create a differential WGRM could also help resolve repetitive genomic structures and extrachromosomal elements with variable GC content and cut patterns. Currently the analysis software (MapSolver) is not publically available. Alternative tools can be developed by labs

with bioinformatics expertise to manipulate the WGM data [35]. We expect that as the WGM technology becomes more accessible and accepted as the norm for sequence assembly QC, free software will become available for researchers.

In addition to the clinical pathogen described in this report, we have incorporated WGM data in the genome sequencing of other clinically relevant isolates of various species, including methicillin resistant *Staphylococcus aureus*, *Staphylococcus haemolyticus*, *Acinetobacter baumannii*, *Klebsiella*, non-equi *Rhodococcus* species, and a putative novel *Bartonella* pathogenic species. With improved integrative NGS-WGM scaffold-building software, the strategy presented here may provide high resolution microbial genotyping data useful for diagnosis and pathogen discovery in clinical settings.

Supporting Information

Table S1 Summary of *de novo* contigs. (XLSX)

Acknowledgments

We thank Erin Newburn and Trevor Wagner of OpGen Inc. for technical support.

Disclaimers

The findings and opinions expressed herein belong to the authors and do not necessarily reflect the official views of the WRAIR, the U.S. Army, or the Department of Defense.

Author Contributions

Discussed results and revised the paper: RJC MCR RAK PEW EPL. Conceived and designed the experiments: JH PEW EPL. Performed the experiments: FOL JH YY MCR. Analyzed the data: FOL JH RJC MCR. Contributed reagents/materials/analysis tools: RAK PEW EPL. Wrote the paper: FOL JH.

References

- Hu B, Xie G, Lo CC, Starkenburg SR, Chain PS (2011) Pathogen comparative genomics in the next-generation sequencing era: genome alignments, pangenomics and metagenomics. *Brief Funct Genomics* 10: 322–333.
- Welch JS, Link DC (2011) Genomics of AML: clinical applications of next-generation sequencing. *Hematology Am Soc Hematol Educ Program* 2011: 30–35.
- Metzker ML (2010) Applications of Next-Generation Sequencing Sequencing Technologies - the Next Generation. *Nature Reviews Genetics* 11: 31–46.
- Nagarajan N, Cook C, Di Bonaventura M, Ge H, Richards A, et al. (2010) Finishing genomes with limited resources: lessons from an ensemble of microbial genomes. *BMC Genomics* 11: 242.
- Loman NJ, Misra RV, Dallman TJ, Constantinidou C, Gharbia SE, et al. (2012) Performance comparison of benchtop high-throughput sequencing platforms (vol 30, pg 434, 2012). *Nat Biotechnol* 30: 562–562.
- Narzisi G, Mishra B (2011) Comparing De Novo Genome Assembly: The Long and Short of It. *PLoS One* 6: e19175.
- Chain PS, Grafham DV, Fulton RS, Fitzgerald MG, Hostetler J, et al. (2009) Genomics. Genome project standards in a new era of sequencing. *Science* 326: 236–237.
- Hongoh Y, Toyoda A (2011) Whole-genome sequencing of unculturable bacterium using whole-genome amplification. *Methods Mol Biol* 733: 25–33.
- Thomas T, Gilbert J, Meyer F (2012) Metagenomics - a guide from sampling to data analysis. *Microb Inform Exp* 2: 3.
- Koser CU, Holden MT, Ellington MJ, Cartwright EJ, Brown NM, et al. (2012) Rapid whole-genome sequencing for investigation of a neonatal MRSA outbreak. *N Engl J Med* 366: 2267–2275.
- Petrosino JF, Highlander S, Luna RA, Gibbs RA, Versalovic J (2009) Metagenomic Pyrosequencing and Microbial Identification. *Clinical Chemistry* 55: 856–866.
- Zhang X, Davenport KW, Gu W, Daligault HE, Munk AC, et al. (2012) Improving genome assemblies by sequencing PCR products with PacBio. *Biotechniques* 53: 61–62.
- Neely RK, Deen J, Hofkens J (2011) Optical mapping of DNA: single-molecule-based methods for mapping genomes. *Biopolymers* 95: 298–311.
- Ramirez MS, Adams MD, Bonomo RA, Centron D, Tolmasek ME (2011) Genomic analysis of *Acinetobacter baumannii* A118 by comparison of optical maps: identification of structures related to its susceptibility phenotype. *Antimicrob Agents Chemother* 55: 1520–1526.
- Schwan WR, Briska A, Stahl B, Wagner TK, Zentz E, et al. (2010) Use of optical mapping to sort uropathogenic *Escherichia coli* strains into distinct subgroups. *Microbiology* 156: 2124–2135.
- Hasan NA, Choi SY, Eppinger M, Clark PW, Chen A, et al. (2012) Genomic diversity of 2010 Haitian cholera outbreak strains. *Proc Natl Acad Sci U S A* 109:E2010–7.
- Jackson SA, Kotewicz ML, Patel IR, Lacher DW, Gangiredla J, et al. (2012) Rapid genomic-scale analysis of *Escherichia coli* O104:H4 by using high-resolution alternative methods to next-generation sequencing. *Appl Environ Microbiol* 78: 1601–1605.
- Lefebvre T, Richards VP, Lang P, Pavinski-Bitar P, Stanhope MJ (2012) Gene Repertoire Evolution of *Streptococcus pyogenes* Inferred from Phylogenomic Analysis with *Streptococcus canis* and *Streptococcus dysgalactiae*. *PLoS One* 7: e37607.
- Lavigne JP, Vergunst AC, Goret L, Sotto A, Combescuré C, et al. (2012) Virulence potential and genomic mapping of the worldwide clone *Escherichia coli* ST131. *PLoS One* 7: e34294.
- Shukla SK, Pantrangi M, Stahl B, Briska AM, Stemper ME, et al. (2012) Comparative Whole-Genome Mapping To Determine *Staphylococcus aureus* Genome Size, Virulence Motifs, and Clonality. *J Clin Microbiol* 50: 3526–3533.
- Ahmed SA, Awosika J, Baldwin C, Bishop-Lilly KA, Biswas B, et al. (2012) Genomic Comparison of *Escherichia coli* O104:H4 Isolates from 2009 and 2011 Reveals Plasmid, and Prophage Heterogeneity, Including Shiga Toxin Encoding Phage stx2. *PLoS One* 7: e48228.
- Fey PD, Iwen PC, Zentz EB, Briska AM, Henkhaus JK, et al. (2012) Assessment of whole-genome mapping in a well-defined outbreak of *Salmonella enterica* serotype Saintpaul. *J Clin Microbiol* 50: 3063–3065.
- Clarridge JE 3rd, Harrington AT, Roberts MC, Soge OO, Maquelin K (2013) Impact of strain typing method on assessing the relationship of paired nares and wound isolates of methicillin-resistant *Staphylococcus aureus*. *J Clin Microbiol* 51:224–31.
- Saunders MP, Wu G, Abuoun M, Pan Z, Anjum M, et al. (2010) Optical genetic mapping defines regions of chromosomal variation in serovars of *S. enterica* subsp. *enterica* of concern for human and animal health. *Epidemiol Infect*: 1–10.
- Shukla SK, Kislow J, Briska A, Henkhaus J, Dykes C (2009) Optical mapping reveals a large genetic inversion between two methicillin-resistant *Staphylococcus aureus* strains. *J Bacteriol* 191: 5717–5723.
- Turner PC, Yomano LP, Jarboe LR, York SW, Baggett CL, et al. (2012) Optical mapping and sequencing of the *Escherichia coli* KO11 genome reveal extensive chromosomal rearrangements, and multiple tandem copies of the *Zymomonas mobilis* *pdc* and *adhB* genes. *J Ind Microbiol Biotechnol* 39:629–39.
- Giongo A, Tyler HL, Zipperer UN, Triplett EW (2010) Two genome sequences of the same bacterial strain, *Gluconacetobacter diazotrophicus* PAI 5, suggest a new standard in genome sequence submission. *Stand Genomic Sci* 2: 309–317.
- Zhou S, Bechner MC, Place M, Churas CP, Pape L, et al. (2007) Validation of rice genome sequence by optical mapping. *BMC Genomics* 8: 278.
- Yi H, Cho YJ, Yong D, Chun J (2012) Genome Sequence of *Escherichia coli* J53, a Reference Strain for Genetic Studies. *J Bacteriol* 194: 3742–3743.
- Triscari-Barberi T, Simone D, Calabrese FM, Attimonelli M, Hahn KR, et al. (2012) Genome sequence of the polychlorinated-biphenyl degrader *Pseudomonas pseudoalcaligenes* KF707. *J Bacteriol* 194: 4426–4427.
- Chandry PS, Gladman S, Moore SC, Seemann T, Crandall KA, et al. (2012) A Genomic Island in *Salmonella enterica* ssp. *salamae* provides new insights on the genealogy of the locus of enterocyte effacement. *PLoS One* 7: e41615.
- Tompkins TA, Barreau G, Broadbent JR (2012) Complete Genome Sequence of *Lactobacillus helveticus* R0052, a Commercial Probiotic Strain. *J Bacteriol* 194: 6349.
- Park J, Zhang Y, Buboltz AM, Zhang X, Schuster SC, et al. (2012) Comparative genomics of the classical *Bordetella* subspecies: the evolution and exchange of virulence-associated diversity amongst closely related pathogens. *BMC Genomics* 13: 545.
- Yoo M, Kim D, Choi KY, Chae JC, Zylstra GJ, et al. (2012) Draft genome sequence and comparative analysis of the superb aromatic-hydrocarbon degrader *Rhodococcus* sp. strain DK17. *J Bacteriol* 194: 4440.
- Nagarajan N, Read TD, Pop M (2008) Scaffolding and validation of bacterial genome assemblies using optical restriction maps. *Bioinformatics* 24: 1229–1235.
- Clifford RJ, Hang J, Riley MC, Onmus-Leone F, Kuschner RA, et al. (2012) Complete Genome Sequence of *Providencia stuartii* Clinical Isolate MRSN 2154. *J Bacteriol* 194: 3736–3737.
- McGann P, Hang J, Clifford RJ, Yang Y, Kwak YI, et al. (2012) Complete sequence of a novel 178-kilobase plasmid carrying *bla*(NDM-1) in a *Providencia stuartii* strain isolated in Afghanistan. *Antimicrob Agents Chemother* 56: 1673–1679.
- Storey S, McGann PT, Lesho EP, Waterman PE (2011) Notes from the Field: Detection of *bla*(NDM-1) Carbapenem Resistance in a Clinical Isolate of *Providencia stuartii* in a U.S./Coalition Medical Facility - Afghanistan, 2011. *Morbidity and Mortality Weekly Report* 60: 756.
- Reslewic S, Zhou S, Place M, Zhang Y, Briska A, et al. (2005) Whole-genome shotgun optical mapping of *Rhodospirillum rubrum*. *Appl Environ Microbiol* 71: 5511–5522.
- Geneious v5.4, Available : <http://www.geneious.com/>. Accessed 2013 Mar 19.
- Riley MC, Kirkup BC Jr, Johnson JD, Lesho EP, Ockenhouse CF (2011) Rapid whole genome optical mapping of *Plasmodium falciparum*. *Malar J* 10: 252.
- Lin J, Qi R, Aston C, Jing J, Anantharaman TS, et al. (1999) Whole-genome shotgun optical mapping of *Deinococcus radiodurans*. *Science* 285: 1558–1562.
- Poirel L, Al Maskari Z, Al Rashdi F, Bernabeu S, Nordmann P (2011) NDM-1-producing *Klebsiella pneumoniae* isolated in the Sultanate of Oman. *J Antimicrob Chemother* 66: 304–306.
- Walsh TR, Weeks J, Livermore DM, Toleman MA (2011) Dissemination of NDM-1 positive bacteria in the New Delhi environment and its implications for human health: an environmental point prevalence study. *Lancet Infect Dis* 11: 355–362.
- Poirel L, Dortet L, Bernabeu S, Nordmann P (2011) Genetic features of *bla*(NDM-1)-positive *Enterobacteriaceae*. *Antimicrob Agents Chemother* 55: 5403–5407.
- Garcia-Martinez J, Acinas SG, Anton AI, Rodriguez-Valera F (1999) Use of the 16S-23S ribosomal genes spacer region in studies of prokaryotic diversity. *J Microbiol Methods* 36: 55–64.
- Chaudhuri RR, Loman NJ, Snyder LA, Bailey CM, Stekel DJ, et al. (2008) xBASE2: a comprehensive resource for comparative bacterial genomics. *Nucleic Acids Res* 36: D543–546.
- Riley MC, Lee JE, Lesho E, Kirkup BC Jr (2011) Optically mapping multiple bacterial genomes simultaneously in a single run. *PLoS One* 6: e27085.