Reducing assembly complexity of microbial genomes with single-molecule sequencing

Sergey Koren1*, Gregory P Harhay2, Timothy PL Smith2, James L Bono2, Dayna M Harhay2, Scott D Mcvey3, Diana Radune1, Nicholas H Bergman1 and Adam M Phillippy1

Abstract

Background: The short reads output by first- and second-generation DNA sequencing instruments cannot completely reconstruct microbial chromosomes. Therefore, most genomes have been left unfinished due to the significant resources required to manually close gaps in draft assemblies. Third-generation, single-molecule sequencing addresses this problem by greatly increasing sequencing read length, which simplifies the assembly problem.

Results: To measure the benefit of single-molecule sequencing on microbial genome assembly, we sequenced and assembled the genomes of six bacteria and analyzed the repeat complexity of 2,267 complete bacteria and archaea. Our results indicate that the majority of known bacterial and archaeal genomes can be assembled without gaps, at finished-grade quality, using a single PacBio RS sequencing library. These single-library assemblies are also more accurate than typical short-read assemblies and hybrid assemblies of short and long reads.

Conclusions: Automated assembly of long, single-molecule sequencing data reduces the cost of microbial finishing to $1,000 for most genomes, and future advances in this technology are expected to drive the cost lower. This is expected to increase the number of completed genomes, improve the quality of microbial genome databases, and enable high-fidelity, population-scale studies of pan-genomes and chromosomal organization.

Background

As the cost of sequencing has dropped, the number of sequencing projects available in the GOLD database [1] has increased 4-fold from 2,905 in 2007 to 11,472 in 2011 [2]. However, many available genomes are heavily fragmented into hundreds or thousands of contigs, and many more are sequenced at low coverage and never submitted. This is in stark contrast to the era before the 'next-generation' revolution, when many genomes underwent expensive manual gap-closing and sequence verification (finishing) before submission [3]. As sequencing costs have dropped, finishing has become impractical given the volume of sequencing data and manual effort required [4]. Only 32% of the genomes in the GOLD database are ‘complete’ or ‘closed’, meaning they contain no gaps. An even fewer number were ‘finished’ by manually correcting errors and adding annotation [5].

This has hampered large-scale, structural analyses of bacterial genomes, and focused research instead on isolated genes and single-nucleotide polymorphisms (SNPs). While it remains impractical to manually finish all but the most important reference genomes, it is now possible to close microbial genomes at a reasonable cost using long-read, single-molecule sequencing and new assembly techniques [6-8]. This is expected to revitalize large-scale comparative genomic studies of whole genomes.

Single-molecule sequencing is a challenging problem that has not, until recently, resulted in a commercial product. Released in 2011, the PacBio RS is the only long-read, single-molecule sequencer currently available. In contrast to competing nanopore approaches [9-11], the PacBio RS utilizes an anchored polymerase and zero-mode waveguide to observe DNA polymerization in real time [12]. This instrument debuted as a rapid method for sequencing outbreak genomes [13,14] and has since been demonstrated on eukaryotic genomes and transcriptomes [8]. Recent studies have focused on

* Correspondence: korens@nbacc.net
1 National Biodefense Analysis and Countermeasures Center, 110 Thomas Johnson Drive, Frederick, MD 21702, USA
Full list of author information is available at the end of the article

© 2013 Koren et al; licensee BioMed Central Ltd. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.
identifying DNA modification, such as methylation patterns, directly from the single-molecule sequencing data [15]. While adoption of this technology was initially slowed by the low accuracy of the single-pass sequences, recent advancements have demonstrated that this drawback can be algorithmically managed to produce assemblies of unmatched continuity [7,8,16]. Steady improvements to the PacBio technology continue to increase read lengths and yield [17], while future technologies promise to combine accuracy with length using either nanopores [11] or advanced sample preparation [18]. Improved microbial genome assembly is an obvious application of these recent developments in long-read sequencing.

Genome assembly is the process of reconstructing a genome from many shorter sequencing reads [19-21]. It is typically formulated as finding a traversal of a properly defined graph of reads, with the ultimate goal of reconstructing the original genome as faithfully as possible. Repeated sequence in the genome induces complexity in the graph and poses the greatest challenge to all assembly algorithms [22]. In addition, repeats are often the focus of analysis [23-25], making their correct assembly critical for subsequent studies. However, repeats can only be resolved by a spanning read or read pair that is uniquely anchored on both sides. Read pairs are typically used due to their length potential (tens of kilobase pairs), but introduce additional complexity because they cannot be precisely sized. Alternatively, long-read sequencing promises to more accurately resolve repeats and directly assemble genomes into their constituent replicons. Figure 1 shows the benefit of increasing read length when assembling Escherichia coli K12 MG1655. This genome can only be assembled into a single contig when the read length exceeds the size of the longest repeat in the genome, a multi-copy rDNA operon. The rDNA operon, sized around 5 to 7 kbp, is the largest repeat class in most bacteria and archaea [26]. Therefore, sequencing reads longer than the rDNA operon, such as those produced by single-molecule sequencing, can automatically close most microbial genomes.

ALLPATHS-LG was the first assembler shown to produce complete microbial genomes using single-molecule sequences [7]. Utilizing a combination of PacBio RS single-molecule reads (2 to 3 kbp), short-range Illumina read pairs (<300 bp insert), and long-range Illumina read pairs (3 to 10 kbp insert), ALLPATHS-LG assembles the Illumina reads first using a de Bruijn graph and incorporates PacBio reads afterwards to patch coverage gaps and resolve repeats. Riberio et al. [7] tested this method on 16 genomes and consensus accuracy was measured at 99.99999% on 3 genomes with an available reference. Four of the sixteen genomes were successfully assembled into a complete genome - the remaining genomes were all highly continuous but left unresolved due to large-scale repeats. These results are promising, especially in terms of consensus accuracy; however, the method requires two different sequencing platforms and three library preparations, which limits its efficiency. In addition, the jumping libraries were observed to be inconsistent at spanning large repeats due to biases in the library construction process.

Ideally, complete genomes could be reconstructed from a single fragment library, minimizing costs. Previously, pair libraries were the only sequencing method capable of spanning large repeats, such as the rDNA operon, but the PacBio RS is now capable of producing single-molecule reads of the same length. Leveraging this recent development, we present an approach for microbial genome closure that relies on overlapping and assembling single-molecule reads de novo rather than
patching and resolving a short-read de Bruijn graph. This exploits the log-normal sequence length distribution of the PacBio RS, which produces a significant fraction of sequences greater than 7 kbp [8]. These long reads can be utilized to span the longest repeat found in most microbial genomes, while the total coverage of reads can be used to construct a high-quality consensus. We estimate that this approach could automatically close >70% of the complete bacteria and archaia in GenBank, without the need for pair libraries, using the currently available PacBio ‘C2’ chemistry. These single-library assemblies are also more accurate than typical short-read assemblies and hybrid assemblies of short and long reads. Finally, we show that the increased sequencing length of future technologies both decreases the coverage requirement and increases the number of genomes that can be closed using this method.

Results and discussion

Early single-molecule sequencing reads were too short and inaccurate to directly perform de novo assembly. Instead, it was shown that the reads could be corrected using a complementary technology prior to assembly [8,14]. However, single-molecule read lengths have continued to improve, from a median length of 747 bp in 2011 to 3,122 bp in 2012 (Figure 2). Due to the increase in length, it is now possible to perform self-alignment and correction. This is because there are more detectable alignment seeds in a longer sequence versus a shorter sequence with equivalent error rate (Figure 3) [28,29]. For example, 1.5 kbp sequences at 10% error are sufficient to reliably detect overlaps, but at 15% error, such as for XL-XL chemistry [30], 3.5 kbp sequences are required [28]. Based on this analysis, and the improving read length of the PacBio RS, we adapted the Celera...
Assembler PBcR pipeline [8] to support correction and assembly using only continuous long reads (CLRs). This new version uses the BLASR software [28] to detect noisy overlaps between reads; an improved version of the PBcR algorithm to process overlaps and correct the long reads; and the Celera Assembler [31] for final assembly. The pipeline is designed to be compatible with future long-read data and has been tested on reads up to 64 kbp in length. The related HGAP software provided by PacBio [32] is a derivative of our correction and assembly pipeline [8] that also performs self-correction of CLR sequences followed by assembly with Celera Assembler. However, HGAP cannot use secondary sequencing data for correction. For consistent comparison to hybrid assembly, all reported results are from the PBcR version of the pipeline.

To evaluate the potential of long-read data to improve microbial genome assembly, we first report the repeat complexity of all complete microbial genomes and predict the fraction that could be closed using a single PacBio sequencing library. We conclude with an analysis of six genomes sequenced using Illumina MiSeq, Roche 454, PacBio circular consensus sequencing (CCS), and PacBio CLR to evaluate performance on real data and compare PacBio CLR self-assembly versus a hybrid approach.

Assembly complexity of NCBI genomes
To describe the complexity of microbial genome assembly using long reads, we define three classes of microbial genomes in terms of repeat content and type, using the common rDNA operon as a benchmark (Figure 4). Class I genomes are defined as genomes with few repeats other than the rDNA operon. Class II genomes contain many mid-scale repeats, such as insertion sequences and simple sequence repeats, but the rDNA operon remains the largest. Class III genomes contain large phage-mediated repeats, segmental duplications, or large tandem arrays that are significantly larger than the rDNA operon. To delineate these classes, all maximal repeats greater than 500 bp and 95% identity were identified for 2,267 finished microbial genomes available from NCBI. Figure 5 illustrates the distribution of maximum repeat size and repeat count for these genomes. The rDNA operon is clearly the largest repeat in most bacteria and archaea, with a tightly bounded size between 5 and 7 kbp. Thus, the threshold for class III is set to a maximum repeat size of greater than 7 kbp. The boundary between classes I and II is less clear, and is set to 100 repeat copies for convenience.

Of the 2,267 genomes analyzed, 69.07%, 7.59%, and 23.33% comprise class I, II, and III, respectively. It is important to note that class I genomes can be assembled well, though not closed, by short-read sequencing, but class II genomes, such as Yersinia pestis, have previously been considered the most difficult to assemble [27]. Now, with single-molecule sequencing reads in excess of 7 kbp, both the mid-range and rDNA repeats can be reliably spanned. This predicts automated closure of class I and II genomes is now possible, and all but the longest class III repeats can be resolved. We note that this analysis is database dependent and may underestimate the true membership of class II and III, because repetitive genomes are the least likely to be complete. A table of repeat count and maximum repeat size for all complete genomes is provided as Additional file 1.

To evaluate the impact of long reads on microbial genome assembly, we simulated error-free sequences following PacBio C1, C2, XL-C2, XL-XL, and projected ‘ZL’ corrected read length distributions. The hypothetical ‘ZL’ chemistry is an extrapolation of annual PacBio chemistry improvements, and is based on a log-normal distribution with double the mean sequence length of the XL-XL chemistry. Sequencing coverage was generated from 50 to 200× for the same 2,267 genomes. For each genome, repeats were considered resolved if at least
one simulated read spanned the entire repeat with unique anchors on both sides. In an overlap-based assembler, this is typically sufficient for resolving a repeat [34]. This estimates the potential resolving power of long reads in the absence of sequencing error. However, since corrected PacBio sequences achieve 99.9% accuracy in practice [8], it is also a reasonable approximation of true data.

Figure 6 shows the predicted number of genomes closed as well as the average number of assembly gaps for all variants of the simulated PacBio reads. For the C2 chemistry, 72.96% of the genomes contain zero gaps at 200× coverage, and the remaining genomes are well assembled but contain large, unresolved repeats. The expected number of gaps is 0.26 ± 3.90, 0.34 ± 1.39, and 2.89 ± 2.92 for class I, II, and III genomes, respectively. The benefit of additional C2 sequencing beyond 100× decreases rapidly, with almost no increase in the number of resolved genomes from 150× to 200× (34 additional genomes). Using the newer XL-C2 chemistry, the number of closed genomes plateaus much earlier, due to the larger number of long sequences. However, at least 50× is still required to generate an accurate consensus using only long-read sequencing [32]. At 200×, upgrading to XL-C2 from C2 chemistry closes an additional 3.79% of genomes; upgrading from XL-C2 to XL-XL closes an additional 11.69%; and upgrading to ‘ZL’ from XL-XL closes an additional 7%. There are diminishing returns in increasing read lengths, because many of the remaining unresolved repeats are more than double the average XL-XL lengths, requiring a jump in average read length to hundreds of kilobases to resolve.

No significant correlation between genome size and assembly continuity was found, which agrees with previous work that found no strong correlation between microbial genome size and repeat coverage [26]. However, assembly complexity is largely influenced by species-specific repeat structure. For example, for 200× C2, the

---

**Figure 4** Three classes of microbial genome assembly complexity. The top row illustrates repeat content via an alignment dotplot in Bacillus anthracis Ames, Yersinia pestis CO92, and Escherichia coli O26:H11 11368. For a repeat occurring at two distinct positions x and y in the genome, a dot of the corresponding size is placed on the matrix at [xy]. The bottom row illustrates assemblies of these genomes using 200× simulated PacBio C2 sequencing (outer circle), and infinite coverage of 500 bp, perfect reads (inner circle). The number of gaps in each assembly is noted. Class I genomes have few repeats except for the rDNA operon sized 5 to 7 kbp. In this case, both short reads and PacBio reads can generate a continuous assembly. Class II genomes have many repeats, such as insertion sequence elements, but none greater than 7 kbp. In this case, the PacBio reads can completely assemble the genome, while the short-read assembly is heavily fragmented. Class III genomes contain large, often phage-related, repeats >7 kbp. In this case, no technology can generate a complete genome. However, the PacBio assembly is significantly more continuous than the short-read assembly.
Figure 5 Repeat count versus maximum repeat length for 2,267 complete genomes. For each genome, the number of repeat regions > 500 bp is given on the horizontal axis and the size of the largest repeat in the genome is given on the vertical axis. A smoothed scatterplot of all complete genomes is in the center, with the corresponding histograms for each axis at the top and right. The figure is cropped to show only repeat counts < 300 and maximum repeat size < 30 kbp. This comprises 95% of the data, with the remaining 5% containing a maximum repeat > 30 kbp or more than 300 repeats. In the extremes, class II genomes can reach over 800 repeat copies, and class III genome repeats can exceed 100 kbp [26,33].

Figure 6 Assembly improvement with increasing coverage and read length. Simulated assembly results on all complete NCBI references as of January 2013 using PacBio RS C1, C2, XL-C2, XL-XL, and projected ZL chemistries. The two figures show the percentage of genomes closed (left) and the average number of remaining gaps (right) with increasing sequencing coverage. C2 and newer chemistries can span the rDNA repeat and thus close many more genomes than the C1 chemistry. However, beyond 150x C2 there is limited benefit from further sequencing because the remaining repeats are too long to resolve. The longer chemistries saturate most repeats and gain little benefit from additional coverage over 50x. Resolving the remaining repeats requires a jump in sequence length to hundreds of kilobases.
expected number of gaps remaining in a Bacillus anthracis or Yersinia pestis strain is 0 ± 0 and 0.42 ± 0.51, respectively. However, some species, such as Escherichia coli (3.04 ± 3.90), exhibit high variance due to strain-specific phage integrations, and so on. An interactive Krona [35] chart detailing the expected number of gaps for all strains, species, genera, and so on under various coverage and chemistry scenarios is available as Additional file 2.

Based on these simulations, 150× is the recommended sequencing depth to maximize assembly continuity using the C2 chemistry. Given sequencing yields of approximately 300 Mbp per SMRT cell after the RS II upgrade [36], this would require three cells for a 5 Mbp genome. Due to the longer XL-C2 chemistry read length, only 100× of XL-C2 is required to maximize closure, or two cells for a 5 Mbp genome. This equates to a lower cost versus previous approaches. The contract sequencing cost for a 5 Mbp genome using the recipe of Riberio et al. [7] is approximately $1,700, and can vary based on library preparation costs and multiplexing efficiencies (Materials and methods). In contrast, the cost of our method, which relies on only a single library preparation, is approximately $1,200 for three SMRT cells of C2, or approximately $900 for two SMRT cells of XL-C2 (Materials and methods). This represents a cost increase versus 100× of Illumina, which can be contracted for $300 or less (Materials and methods), but the resulting Illumina assemblies are typically in hundreds of contigs and require heavy multiplexing to minimize costs [37,38].

Assembly and closure of real data

To validate our approach, we sequenced the genomes of six bacteria of varying complexity and GC content (Table 1). For each genome, at least 200× of PacBio C2 CLRs [12] were generated, along with 454 FLX + [39], Illumina MiSeq [40], and PacBio CCS for comparison to short-read and hybrid assemblies. For uniformity, datasets exceeding 200× CLR, 50× 454, 100× Illumina MiSeq, or 25X CCS were randomly down-sampled to these limits to reflect typical coverage depths. For the five novel genomes, a closely related genome was used to estimate assembly complexity. In the case of E. coli K12, the available reference sequence was used (Materials and methods).

For long-read assembly, the genomes were then assembled using PacBio CLR in isolation and CLR corrected with the secondary technologies. The assemblies largely matched the simulated expectation, independent of the approach used (Table 2; chi-squared P-value <0.015). The class I genomes E. coli K12, Bibersteinia trehalosi, and Salmonella enterica Newport, and the borderline class III genome Mannheimia haemolytica were all brought to closure using self-correction and at least one of the technology combinations. In nearly all cases, the assemblies of self-corrected CLR sequences outperformed the hybrid assemblies in terms of continuity, error rate, and the assembly likelihood score. In contrast, assemblies of the hybrid data showed greater variability in performance, likely due to subtle errors introduced during correction. Only the self-corrected CLR assembly of F. tularensis did not match the simulated expectation. However, it is noted that this dataset has the lowest mean and median sequence length as well as a low fraction of sequences longer than 7 kbp relative to the other projects (8.52% versus an average of 13.47%). A machine failure occurred towards the end F. tularensis sequencing, possibly explaining the reduced performance of the preceding cells. An additional

| Table 1 Total sequence and estimated coverage for six sequenced genomes |
|--------------------------|-----------------|-----------------|-----------------|------------------|-----------------|-----------------|-----------------|-----------------|
| Organism                | Genome Mbp      | GC % Class      | PB CLR Gbp      | PB % error (200×) | PB % >7 kbp (200×) | PB coverage >7 kbp (200×) | PB CCS Gbp      | 454 Gbp         | MiSeq Gbp       |
| Escherichia coli K12 MG1655 | 4.65            | 50% I           | 1.4 (294×)      | 10.44            | 16.07            | 32.11×           | 0.2 (44×)       | 0.19 (42×)      | 1.73 (372×)      |
| Escherichia coli O157:H7 F8092B [41] | 5.52            | 50% III         | 2.2 (397×)      | 10.49            | 13.37            | 26.75×           | NA             | 0.22 (46×)      | 0.65 (118×)      |
| Bibersteinia trehalosi USDA-ARS-USMARC-192 | 2.41            | 41% I           | 1.1 (439×)      | 13.17            | 5.56             | 11.12×           | 0.09 (35×)      | 0.15 (59×)      | 0.62 (247×)      |
| Mannheimia haemolytica USDA-ARS-USMARC-2286 | 2.55            | 41% III         | 0.5 (212×)      | 10.46            | 23.06            | 47.12×           | 0.1 (42×)       | NA             | 0.39 (85×)       |
| Francisella tularensis 99A-2628 | 1.90            | 32% II          | 0.7 (331×)      | 11.82            | 8.52             | 17.03×           | NA             | 0.09 (40×)      | 11.73 (5870×)    |
| Salmonella enterica Newport SN31241 [42] | 5.01            | 52% I           | 1.1 (217×)      | 12.33            | 9.30             | 18.61×           | 0.1 (22×)       | 0.13 (25×)      | 0.28 (56×)       |

The table lists available sequencing coverage (in base pairs and estimated fold coverage) of six bacterial genomes chosen to validate the closure recipe, for the four types of data produced (PacBio CLR, PacBio CCS, 454 titanium, and MiSeq paired end). Sequencing data were randomly down-sampled following the procedure described in Materials and methods. For each genome, the expected genome size, GC content, and complexity class is included. PB % error (200×): the estimated error rate based on BLASR [28] mappings of the analyzed random 200× subset to the assembly. PB % >7 kbp (200×): the percentage of bases in sequences longer than 7 kbp, before correction. PB coverage >7 kbp (200×): the coverage represented by sequences longer than 7 kbp, before correction. NA, not available; PB, PacBio.
100× coverage was generated for *F. tularensis*, bringing the assembly to expectation (Table 2).

For comparison to short-read sequencing and assembly, the 454 and MiSeq reads were assembled in isolation and the results compared to PacBio CLR assemblies (Table 3). As of writing this manuscript, the estimated sequencing costs for the assemblies presented in Table 3 are $300 to $500 for 100× Illumina (paired HiSeq 2500 or MiSeq), $4,700 for 50× 454 (unpaired FLX+), and $1,400 for 200× PacBio (RS II C2). These estimates include library preparation and assume multiplexing efficiencies (Materials and methods). Compared to the PacBio assemblies, the short-read assemblies were significantly less continuous, with well over 100 gaps and a 30-fold reduction in N50 contig size, on average. The cost to manually close these assemblies - estimated by Riberio *et al.* [7] to exceed $13,000 - is an order of magnitude higher than any of the single-molecule methods (Materials and methods). These results are consistent with expectation based on the short read lengths and repeat complexity of the presented genomes.

### Table 2 Genome assembly continuity and correctness using hybrid and self-correction approaches

| Organism                  | Corrected by          | Assembly bp | Number of contigs (expected) | Number of contigs (actual) | N50 (expected) | N50 (actual) | LAP | Number of discordant bases | QV |
|---------------------------|-----------------------|-------------|------------------------------|----------------------------|----------------|--------------|-----|----------------------------|----|
| *E. coli* K12             | Reference             | 4,639,675   | 1                            | 4,639,675                  | NA             | −9.65E + 07  | 4   | >60                        |    |
| MiSeq 100x                | 4,647,253             | 1           | 2                            | 3,089,011                  | −9.64E + 07    | 4            | >60                        |    |
| 454 50x                   | 4,649,004             | 1           | 1                            | 4,649,004                  | −9.64E + 07    | 3            | >60                        |    |
| CCS 25X                   | 4,653,267             | 1           | 1                            | 4,653,267                  | −9.64E + 07    | 3            | >60                        |    |
| **Self**                  | 4,653,486             | 1           | 1                            | 4,653,486                  | −9.64E + 07    | 3            | >60                        |    |
| *E. coli* O157:H7         | Near neighbor         | 5,594,477   | 3                            | 3,776,951                  | NA             | −3.82E + 07  | 1,282 | 36.40                      |    |
| MiSeq 100x                | 5,624,394             | 10          | 12                           | 927,294                    | −3.67E + 07    | 13           | 56.35                      |    |
| 454 40x                   | 5,613,057             | 10          | 12                           | 927,294                    | −3.67E + 07    | 13           | 56.35                      |    |
| **Self**                  | 5,611,389             | 10          | 9                            | 4,324,437                  | −3.66E + 07    | 0            | >60                        |    |
| *B. trehalosi*            | MiSeq 100x            | 2,402,545   | 6                            | 1,603,511                  | −3.28E + 07    | 1            | >60                        |    |
| 454 50x                   | 2,413,761             | 4           | 1                             | 2,411,501                  | −3.27E + 07    | 0            | >60                        |    |
| CCS 25X                   | 2,411,501             | 1           | 2                             | 2,411,501                  | −3.27E + 07    | 0            | >60                        |    |
| **Self**                  | 2,411,068             | 1           | 2                             | 2,411,068                  | −3.27E + 07    | 0            | >60                        |    |
| *M. haemolytica*          | MiSeq 100x            | 2,712,467   | 1                             | 2,712,467                  | −3.31E + 07    | 0            | >60                        |    |
| CCS 25X                   | 2,739,949             | 2           | 2                             | 2,686,992                  | −3.31E + 07    | 0            | >60                        |    |
| **Self**                  | 2,736,037             | 1           | 2                             | 2,736,037                  | −3.31E + 07    | 0            | >60                        |    |
| *F. tularensis*           | Near neighbor         | 1,895,727   | 1                             | 965,253                    | −1.33E + 07    | 113          | 42.25                      |    |
| MiSeq 100x                | 1,879,071             | 3           | 10                            | 357,518                    | −1.33E + 07    | 0            | >60                        |    |
| 454 50x                   | 1,863,947             | 3           | 15                            | 201,203                    | −1.33E + 07    | 0            | >60                        |    |
| CCS 25X                   | 1,828,135             | 3           | 8                             | 401,731                    | −1.33E + 07    | 0            | >60                        |    |
| **Self**                  | 1,877,407             | 3           | 3                             | 573,021                    | −1.33E + 07    | 0            | >60                        |    |
| *S. enterica* Newport     | Near neighbor         | 5,007,719   | 2                             | 4,827,641                  | −2.26E + 07    | 20           | 53.99                      |    |
| MiSeq 56X                 | 5,027,784             | 4           | 2                             | 4,918,796                  | −2.24E + 07    | 2            | >60                        |    |
| 454 25X                   | 5,034,500             | 4           | 3                             | 4,095,943                  | −2.24E + 07    | 2            | >60                        |    |
| CCS 22X                   | 5,030,885             | 4           | 2                             | 4,921,886                  | −2.24E + 07    | 2            | >60                        |    |
| **Self**                  | 5,029,197             | 4           | 2                             | 4,919,684                  | −2.24E + 07    | 2            | >60                        |    |

**Organism**: the genome being assembled. **Corrected by**: the short-read data used for correction. **Assembly bp**: the total number of base pairs in all contigs (only contigs containing at least 100 reads are included in all results). **Number of contigs (expected)**: predicted number of contigs for a known reference (or near-neighbor). **Number of contigs (actual)**: the number of contigs comprising the assembly. **N50**: N such that 50% of the genome is contained in contigs of length ≥N. **LAP**: the assembly likelihood score. A score closer to zero indicates a better assembly. **Number of discordant bases**: the number of SNPs and indels identified by mapping MiSeq sequences back to the assembly and recording discrepancies. Each incorrect base is counted (that is, an indel that is a deletion of two bases from the assembly counts as two in this column). **QV**: estimated from the number of discordant bases as \( \log_{10} \left( \frac{\text{assembly length}}{\text{incorrect bases}} \right) + 10 \). The QV can be converted to an error probability \( P = 10^{\frac{-QV}{10}} \). Assemblies were generated by Celera Assembler [31] followed by post-processing with Quiver [32]. NA, not available.

### Long-read assembly validation

Reference-free assembly validation was performed on all assemblies using mapped Illumina MiSeq reads to estimate consensus error rates and determine an assembly likelihood score. For the majority of assemblies that did not utilize the MiSeq data, this represents an independent
validation using a complementary data source. Consensus accuracy was estimated from single nucleotide discrepancies identified between the Illumina data and the assembly using FreeBayes [46], and assembly scores were computed using FRCbam [47], ALE [48], and LAP [49]. The latter two measure the likelihood of a set of Illumina reads being generated by the assembly - essentially how consistent the assembly is with the reads. FRCbam, ALE, and LAP scored assemblies similarly, so Tables 2 and 3 list only the LAP scores and estimated consensus accuracy (a full validation report is provided as Additional file 3). In all cases self-correction produced the best LAP score with consistently high accuracies. The self-corrected assemblies averaged 99.9993% accuracy prior to polishing, and >99.99995% after re-alignment and polishing with Quiver. The second-generation assemblies averaged 99.9993% and 99.9992% accuracy for 454 and MiSeq, respectively. For comparison, a consensus accuracy of 99.999% (1 error in 100,000 bp) is considered finishing quality [4]. Thus, the automated assemblies of self-corrected PacBio CLR sequences surpass both the quality of second-generation assemblies and the quality standard for finished genomes.

A finished reference is available for *E. coli* K12 MG1655, which was one of the genomes sequenced here. However, the reference-free validation indicated the Quiver-polished assemblies of *E. coli* K12 were more consistent with the Illumina reads than the MG1655 reference sequence. The assemblies have both a better likelihood score and fewer single-nucleotide mapping discrepancies than the reference, suggesting that most differences between the assemblies and the published reference are true isolate-level variations. Table 4 reports these differences using the GAGE metrics [37].

Assembly performance varied depending on the correction method. On average, 454 correction was less accurate, which is unsurprising given the known homopolymer bias of this technology [39]. Sharing 454's length and error characteristics, a similar result would be expected if correcting with Ion Torrent data [50]. The CCS correction also underperformed the other methods, likely due to its lower per-read accuracy. Most promising, the self-corrected CLR sequences produced the fewest errors, even outperforming assemblies that included the Illumina data used for validation. This is consistent with the PacBio RS having low systematic bias, allowing a high-quality consensus to be generated even from low-accuracy reads [8]. Longer sequences can also be aligned more accurately, provided they are long enough to compensate for the error rate [29]. These results demonstrate that high-quality, high-continuity bacterial assemblies can now be generated using exclusively single-molecule sequencing data.

Table 3 Genome assembly continuity and correctness comparison to secondary technologies

| Organism          | Assembled with | Assembly bp | Contigs | N50 | LAP | Discordant bases | QV  |
|-------------------|----------------|-------------|--------|-----|-----|------------------|-----|
| *E. coli* K12     | MiSeq 100x 2x150 bp 300 bp (MaSuRCA iCORN) | 4,682,345 | 139    | 113,852 | −968E + 07 | 28 | 52.23 |
| 454 50x           | 4,569,757 | 93 | 117,490 | −937E + 07 | 17 | 54.29 |
| **PbCr 200x**     | 4,653,486 | 1 | 4,653,486 | −964E + 07 | 3 | >60 |
| *E. coli* O157:H7 | MiSeq 100x 2x150 bp 500 bp (SPAdes iCORN) | 5,433,737 | 413 | 133,641 | −367E + 07 | 62 | 49.43 |
| 454 22x + 8x 5 kbp + 10x 10 kbp | 5,347,050 | 409 | 133,665 | −373E + 07 | 66 | 49.09 |
| **PbCr 200x**     | 5,611,389 | 9 | 4,324,437 | −366E + 07 | 0 | >60 |
| *B. trehalosi*    | MiSeq 100x 2x150 bp 500 bp (SPAdes iCORN) | 2,377,594 | 83 | 222,446 | −331E + 07 | 10 | 53.76 |
| 454 50x           | 2,364,704 | 66 | 117,742 | −332E + 07 | 9 | 54.20 |
| **PbCr 200x**     | 2,411,068 | 1 | 2,411,068 | −327E + 07 | 0 | >60 |
| *M. haemolytica*  | MiSeq 100x 2x150 bp 500 bp (MaSuRCA iCORN) | 2,721,965 | 89 | 84,094 | −333E + 07 | 47 | 47.63 |
| **PbCr 200x**     | 2,736,037 | 1 | 2,736,037 | −331E + 07 | 0 | >60 |
| *F. tularensis*   | MiSeq 100x 2x250 bp 500 bp (SPAdes iCORN) | 1,825,374 | 130 | 240,655 | −133E + 07 | 0 | >60 |
| 454 50x           | 1,655,657 | 326 | 7,316 | −133E + 07 | 28 | 47.72 |
| **PbCr 300x**     | 1,877,407 | 3 | 573,021 | −133E + 07 | 0 | >60 |
| *S. enterica* Newport | MiSeq 56x 2x150 bp 500 bp (MaSuRCA iCORN) | 5,187,269 | 114 | 195,780 | −224E + 07 | 360 | 41.59 |
| 454 23x + 2x 10 kbp | 5,005,089 | 172 | 372,513 | −225E + 07 | 39 | 51.08 |
| **PbCr 200x**     | 5,029,197 | 2 | 4,919,684 | −224E + 07 | 2 | >60 |

Organism: the genome being assembled. Assembled with: the sequencing data used for assembly. 454 sequencing was unpaired FLX+, with paired-end sequencing available for some genomes, as indicated. MiSeq sequencing was paired-end, indicated as 2xXbp Yb where X is the target read length and Y is the paired-end size. Column definitions are the same as in Table 2. PacBio RS sequences were self-corrected and assembled as in Table 2. 454 sequences were assembled with Newbler [39] and MiSeq sequences were assembled with SPAdes [43] and MaSuRCA [38,44]. Both assemblies were polished using iCORN [45] and the one with the best LAP score was reported.
Future technology

All sequencing experiments above used the PacBio C2 chemistry released in February 2012. More recently, PacBio released the updated XL chemistry. Using *E. coli* K12 XL-C2 sequencing data provided by PacBio, we modeled the corrected read length distribution and simulated error-free sequences for the same 2,267 reference genomes as before. Using the longer sequences, more genomes are closed at lower coverage. This is due to the larger number of sequences over 7 kbp (22% for XL-C2 versus 16% with C2). Trading increasing read length for decreased accuracy can negatively impact alignment and assembly (for example, XL-C2 versus XL-XL; Figure 3). For this reason, actual C2 and XL-C2 reads were found to assemble better than XL-XL in practice (data not shown). The RS II instrument upgrade increased throughput to 300 Mbp per SMRT cell. Based on these numbers, two XL-C2 SMRT cells will be sufficient to close over 70% of known microbial genomes automatically, for less than $1,000 per genome. This includes the vast majority of class I and II genomes, and predicts an average number of gaps of 2.93 ± 2.90 for class III genomes as before. Using the longer sequences, more genomes can be assembled into fewer than 10 contigs using currently available technology. This increase in continuity greatly reduces the required cost of manual closure, which directly correlates with the number of gaps. Complementary closure techniques, such as optical mapping [51], are also enhanced by longer contigs and can be used to efficiently close even the most complex genomes.

Long reads present a great opportunity, but also new challenges. For example, small replicons shorter than the typical 10 kbp library size may be inadvertently excluded from sequencing. We also noted low-abundance structural polymorphism in many of the samples analyzed. These mixed polymorphisms (for example, inversions) would have been easily overlooked in fragmented assemblies of short reads. However, to fully capture such structural dynamics requires a graph-based representation of the genome, such as FastG [52], that allows for allelic diversity. In addition, long reads present algorithmic challenges to existing assemblers. Most existing assemblers are incapable of exploiting long reads. Celera Assembler and MIRA [53] are two exceptions, but these programs were developed for reads no more than 1 kbp and become cumbersome for very long reads. New algorithms, especially for consensus generation via multiple alignment, are needed to extract the full potential from these new data.

Conclusions

Long, single-molecule reads are sufficient for the complete assembly of most known microbial genomes. The assemblies presented here have good likelihood and finished-grade consensus accuracy exceeding 99.9999%. By exploiting a model of the sequencing process, Quiver is able to improve assembly accuracy by QV 10, on average; and while there may be undiscovered biases in single-molecule sequencing, PacBio consensus accuracy always exceeded that of the second-generation sequencing data. In addition, assemblies of only single-molecule data consistently matched or exceeded the quality of both short-read and hybrid assemblies. Lastly, this approach requires only a single sequencing library, and reduces the time and cost of closure compared to previous approaches. However, for applications such as high-throughput SNP typing, draft Illumina sequencing is likely to remain the preferred option due to current throughput and cost advantages.

For class III genomes that cannot be closed using current single-molecule sequencing, assembly continuity is significantly improved over first- and second-generation sequencing. In simulations, these most difficult genomes average only 2.89 ± 2.92 gaps with C2 sequencing, suggesting that 99% of all known microbial genomes can be assembled into fewer than 10 contigs using currently available technology. This increase in continuity greatly reduces the required cost of manual closure, which directly correlates with the number of gaps. Complementary closure techniques, such as optical mapping [51], are also enhanced by longer contigs and can be used to efficiently close even the most complex genomes.
Finally, it is expected that future improvements to the PacBio chemistry, or the release of new technologies, will further extend the reach of microbial genome assembly. For example, the recent median read increase from PacBio C2 to XL chemistries (approximately 2 kbp to approximately 3 kbp) is predicted to reduce the recommended coverage requirement by two-fold. Thus, it is reasonable to expect that future chemistries with increased read lengths, and the corresponding throughput increases, will allow the full closure of most known bacteria and archaea at a cost of well under $1,000 a genome. This cost will continue to fall with future technology advances, improving reference database quality and enabling population-scale research on the structural dynamics of microbial genomes.

Materials and methods
GOLD database and NCBI genomes
To estimate the number of complete versus draft genomes, searches were performed on 4 March 2013 at Genomes OnLine Database (GOLD) [54]. The total number of projects with status ‘Complete and Published’ was 2,427. The total number with status ‘Permanent Draft’ was 1,752. The total number of projects of any status with sequencing status ‘Draft’ was 3,389. A single project had project status ‘Complete and Published’ and sequencing status ‘Draft’ and 7 projects had project status ‘Permanent Draft’ and sequencing status ‘Draft’. These were excluded from the calculations to avoid double counting. The percentage of closed genomes was then computed as:

\[
\frac{2427}{(1752 + 2427 + 3389 - 1 - 7)} = \frac{2427}{7560} = 32.10\%.
\]

Closed bacterial and archaeal genomes were obtained from NCBI [55] on 17 January 2013. This dataset contained 2,245 complete genomes (including constituent plasmids). The data were manually curated to remove eight plasmid-only genomes and associate the loose plasmids with the appropriate genome based on identifiers listed in BioProject. Also, a total of 30 genomes combined more than a single assembly/BioProject and were separated. This resulted in the 2,267 (2,245 - 8 + 30) genomes used for analysis.

Repeat analysis
Genomic repeats were identified using Nucmer [56] nucmer -maxmatch -nosimplify and filtered using the associated delta-filter command delta-filter -1500 -i=95 to retain only repeats greater than 500 bp and over 95% copy identity. Self-alignments on the main diagonal were discarded, and the repeat matches reduced to a set of intervals along the genome. Any interval contained within a larger interval was discarded; repeat count computed as the number of remaining intervals; and the largest interval noted as the maximum repeat size.

For each genome, error-free reads were uniformly sampled at 50 to 200× coverage and the read lengths were randomly chosen, with replacement, from real PBCR sequence distributions (E. coli K12 C1, C2 and XL-C2, and XL-XL chemistries) and a hypothetical future chemistry (‘ZL’). A list of genomic repeats was compiled, as described above, and any abutting or overlapping repeats were merged. The simulated sequences were then mapped back to the genome using nucmer -maxmatch. A repeat was considered resolved if at least one read spanned the full length of the repeat with at least 40 bp uniquely anchored on both sides. This reflects the default minimum overlap length needed by Celera Assembler to resolve a repeat. The simulation returns the expected number of contigs after breaking at any unspanned repeats.

For C1 chemistry, the length distribution was based on previously published results [8] using 50× CLR sequences and 100× Illumina for correction. For C2 chemistry, the length distribution was based on 200× of sequence corrected using 25× CCS. For XL-C2, the length distribution was based on 100× of sequence corrected using 25× CCS. For XL-XL chemistry, the length distribution was based on 50× distribution of match lengths after mapping sequences to the reference. The XL-XL sequences were later corrected and the observed corrected distribution closely matched the mapped distribution (mean 4,104.91 and 4,690.30 respectively, max 27,095 and 25,320, respectively). The ‘ZL’ chemistry was based on a doubling of the XL-XL sequence length, similar to the past increase from C2 (mean = 2,476) to XL-XL (mean = 4,105). Log-normal distributions were fit to the data using the R function rlnorm(100000, mean(log(values)), sd(log(values))) and the maximum was limited to \( \mu + 5 \sigma \). The mean/standard deviation values were (6.69, 0.37), (7.59, 0.67), (7.90, 0.63), and (8.02, 0.79) for C1, C2, XL-C2, and XL-XL distributions, respectively.

PBCR correction pipeline
Two notable improvements were applied to the PBCR algorithm [8]. One to improve detection of SMRTbell adapter sequences, and the other to fill coverage gaps introduced by the correction process. To remove SMRTbell adapters, short reads mapped to each long read are examined.

If multiple short reads match in both a forward and reverse orientation around a common point, the long read is split at this position. To identify sequences with multiple breakpoints, the above procedure can be applied recursively to the split subsequences. To patch an alignment coverage gap, short reads surrounding the gap
are first identified. All other long reads containing these short reads in the same order and orientation are recruited and the consensus of the long read sequences is used to fill the gap. Only the surrounding short reads within a fixed window are used for recruitment. On an *E. coli* K12 test dataset, the average corrected read length increased to 4,187 from 2,493 when this feature was enabled while maintaining a corrected read identity above 99.6%. The assembly N50 also increased from 3.32 Mbp to 4.65 Mbp.

It was previously assumed that correction using only PacBio CLRs was not feasible [8]. However, this analysis was based on the C1 chemistry with a median read length of only 540 bp and error rate of 16.3%. We reproduced the analysis in Chaisson and Tesler [28] on varying read lengths and error rates. Using C1 and C2 sequencing data, we compared the length and identity of the overlaps when mapping sequences to themselves versus mapping them to the reference. The predicted overlaps closely match the expected overlaps for alignments of C2 reads. However, C1 overlaps are under-detected because they are not sufficiently long to compensate for their error rate. Thus, self-correction is not feasible with the C1 chemistry, but more recent chemistries (C2 onward) allow self-correction due to the increased read length.

**Sequence generation**

Libraries were prepared for each bacterial strain using kits provided by the manufacturer of the sequencing platform, as suggested by the product manuals. PacBio CLRs were generated from libraries made with genomic DNA sheared to an average of 8 to 10 kb using either HydroShear (Digilab, Marlborough, MA, USA), or Covaris G-tube (Covaris, Inc., Woburn, MA, USA). SMRTbell libraries were prepared from these fragments, and bound to eC2, C2, or XL versions of polymerase as suggested by the manufacturer. For eC2 and C2 polymerases, bound complexes were passively loaded into the SMRT cells on the instrument. For XL polymerases, bound complexes were adhered to MagBeads as recommended and actively loaded. At the time of this data collection, the stage start option for sequence collection was not available, so the default mode of data collection was used. PacBio CCS reads were generated from libraries made with genomic DNA sheared to an average of 300 to 800 bases using a Covaris S220 instrument according to the instrument recommendations for these fragment sizes. The libraries were bound to C2 polymerase, and passively loaded into the SMRT cells for sequencing.

Illumina libraries were prepared using TruSeq DNA sample prep kits (Illumina Inc., San Diego, CA, USA) as recommended by the included instructions. DNA was sheared to approximately 500 to 800 bases prior to library construction using a Covaris S220 instrument. The libraries were sequenced using 2 × 150 or 2 × 250 paired end protocols on a MiSeq instrument (Illumina), as recommended by the manufacturer.

Libraries for sequencing on the GS FLX + platform (454 Life Sciences, Branford, CT, USA) were prepared with GS Rapid Library Prep Kits as suggested by the manufacturer. Genomic DNA was sheared to approximately 2 kbp prior to library construction using a Covaris S220, and sequenced on the instrument using recommended emPCR and sequencing conditions for GS FLX + sequencing kits.

**Long-read correction and assembly**

For each genome, 200× of PacBio CLR sequences were corrected using pacBioToCA. In addition to self-correction, hybrid correction using 100× MiSeq, 50× 454, and 25× CCS was performed. Whenever more data were available, it was randomly down-sampled to these values for consistency. All runs used the same default parameters with the exception of the genome size, which was manually approximated beforehand for each genome. After correction using CCS, MiSeq, or 454, the corrected sequences were trimmed by quality and subsampled before assembly. The trimming procedure is an automated step that selects, via dynamic programming, the largest range for each corrected read such that the average consensus quality score exceeds QV 54.5. The minimum overlap length was adjusted based on the average length of the trimmed sequences. After trimming, only the longest 25× of the corrected reads were assembled with Celera Assembler, which, while originally designed for Sanger [57] sequences, has since been adapted to 454 [31] and PacBio RS [8] sequences. Overlap-based assemblers tend to underperform on high coverage data [19,21], so this sampling step both reduces runtime and helps improve assembly continuity. After assembly, contigs with fewer than 100 reads were discarded and the rest polished following PacBio’s guidelines. The assembly was imported as the reference and eight SMRT cells (five for FT) were aligned to the genome using the RS_Resequencing pipeline. SMRTAnalysis 1.4.0 was run as smrtpipe.py --params=settings.xml xml::input.xml. Any lower-case bases or ambiguities (Ns) remaining were trimmed from the beginnings and ends of contigs.

**Secondary sequencing correction and assembly**

For each genome, the same data that were used for correcting PacBio RS sequences was also assembled in isolation. At most 50× of 454 data were assembled using Newbler v2.8 [39]. Newbler ran as runAssembly --o asm --cpu 8 *.sff. At most 100× of Illumina MiSeq data were assembled using SPAdes v2.5.0 [43] and
MaSuRCA v1.9.5 [44] as these were the top-performing assemblers in the GAGE-B competition [38]. SPAdes ran as spades.py --k 21,33,55,77 --careful --pe1-1 miseq.1.fastq --pe1-2 miseq.2.fastq --o spades. MaSuRCA ran as runSRCA.pl config.txt followed by bash assemble.sh. The contents of the config.txt file were:

```plaintext
PATHS
JELLYFISH_PATH=/full/path/to/MSR-CA/bin
SR_PATH=/full/path/to/MSR-CA/bin
CA_PATH=/full/path/to/MSR-CA/CA/Linux-amd64/bin
END
DATA
PE=pe 500 200 miseq.1.fastq miseq.2.fastq
END
PARAMETERS
GRAPH_KMER_SIZE=auto
USE_LINKING_MATES=1
LIMIT_JUMP_COVERAGE=60
CA_PARAMETERS=ovlMerSize=30
cgwErrorRate=0.25 ov1Memory=4GB
KMER_COUNT_THRESHOLD=1
NUM_THREADS=16
JF_SIZE=5000000000
DO_HOMOPOLYMER_TRIM=0
END
```

Both SPAdes and MaSuRCA assemblies were polished using iCORN [45]. The 454 assembly along with the four Illumina MiSeq assemblies were validated as described below. Only the best scoring Illumina assembly for each genome is included in Table 3 (a full validation report for all generated assemblies is provided as Additional file 3).

### Validation

For *E. coli* K12, the MG 1655 reference [GenBank: NC000913] is available. Reference-based validation was performed using the GAGE scripts and metrics [37]. The SNP count between references and assemblies was calculated using Nucmer and show-snps [56]. For all genomes, reference-free validation was also performed. When no reference was available, a near neighbor was included for comparison in validation results. For *E. coli* O157:H7 the reference *E. coli* O157:H7 Sakai [GenBank:NC_002127], [GenBank:NC_002128], [GenBank:NC_002695] was used. For *F. tularensis*, the reference *F. tularensis* subsp. Holarctica OSU18 [GenBank:NC_008369] was used. For *S. enterica* Newport, the reference *S. enterica* Newport SL254 [GenBank:NC_011079], [GenBank:NC_011080], [GenBank:NC_009140] was used. For *S. enterica* Newport, Illumina reads and contigs corresponding to two short plasmids (<10 kbp) and the phiX control were removed prior to validation. Reference-free validation was performed with FRC [47], ALE [48], and LAP [49]. These tools require paired-end sequences for validation, so Illumina data were mapped to the assemblies. For FRC, the bowtie command bowtie –I <min> -X <max> –f –l2 5 –e1 4 0 –best –k1 –S was used, based on the example provided with the source. For ALE, the bowtie command bowtie –I <min> -X <max> –f –l1 0 –e300 –a –v1 –S was used. LAP has a built-in bowtie2 procedure that was used. To call SNPs, a random 100x subset of reads was selected from the mapped, left end of Illumina pairs. Left ends were selected as they were observed to be higher quality and a larger fraction mapped to the assemblies. From these reads, SNPs and indels were called using FreeBayes. FreeBayes was run with the command freebayes –C2 –O –q2 0 –F 0.51 –z 0.02 –X –U –p1.

### Sequencing cost estimate

Sequencing costs for PacBio RS and Illumina library preparation were taken on 16 July 2013 from the Duke Genome Sequencing and Analysis Core Resource [58]. The price for an external domestic (USA) institution was used. The throughput of a single PacBio RS cell was assumed to be 125 Mbp for CLR and 20 Mbp for CCS. The throughput of a single PacBio RS II cell was assumed to be 300 Mbp. The throughput of a GS FLX + was assumed to be 700 Mbp. The throughput of a MiSeq run was assumed to be 3 Gbp. The throughput of a HiSeq flow cell was assumed to be 300 Gbp (37.5 Gbp per lane). Library costs are listed as advertised by Duke, but could potentially be lowered for large-scale projects via automation.

The cost of multiplex sequencing on a HiSeq 2500 was computed as:

\[
100X \text{ Illumina Paired-end} = \text{Library prep} + 1/75 \text{ HiSeq lane} \\
= $283.05 + $39.44 \\
= $322.49
\]

The cost of multiplex sequencing on a MiSeq was computed as:

\[
100X \text{ Illumina Paired-end} = \text{Library prep} + 1/6 \text{ MiSeq Run} \\
= $283.05 + $196.55 \\
= $479.60
\]
The cost of half-plate of sequencing on a 454 GS FLX + was computed as:

\[
50 \times 454 \text{ Fragment} = \text{Library prep} + \frac{1}{2} \text{ Plate} \\
= \$834.99 + \$3,839.84 \\
= \$4,674.83
\]

The cost of the ALLPATHS-LG [7] recipe was computed as:

\[
50 \times \text{Illumina Paired-end} = \text{Library prep} \\
+ \frac{1}{12} \text{ MiSeq Run} \\
= \$283.05 + \$98.28 \\
= \$381.33
\]

\[
50 \times \text{Illumina Mate-pair} = \text{Library prep} \\
+ \frac{1}{12} \text{ MiSeq Run} \\
= \$667.03 + \$98.28 \\
= \$765.31
\]

\[
50 \times \text{PacBio RS II} = \text{Library prep} + 1 \text{ SMRT cell} \\
= \$365.05 + \$252.22 = \$617.27
\]

Total = $1,763.91 ($1,628.98 using 96-way multiplexed HiSeq).

This is consistent with the estimate of $1,669 for Illumina sequencing and $1,365 for PacBio RS sequencing (total $3,034) given in the supplementary materials of Riberio et al. [7], adjusted for cost decreases over the past year.

The cost of PacBio RS II sequencing was computed as:

\[
25 \times \text{PacBio RS CCS} = \text{Library prep} + 6 \text{ SMRT cells} \\
= \$365.05 + \$1,513.32 \\
= \$1,878.37
\]

\[
150 \times \text{PacBio RS C2} = \text{Library prep} + 6 \text{ SMRT cells} \\
= \$418.68 + \$1,513.32 \\
= \$1,932.00
\]

Given the recent PacBio RS II update, the per-cell yields have increased from 125 Mbp to 300 Mbp. Thus, two SMRT cells are sufficient for 100× coverage and three SMRT cells are sufficient for 150× coverage. The cost of RS II sequencing was computed as:

\[
200 \times \text{PacBio RS II C2} = \text{Library Prep} + 4 \text{ SMRT cells} \\
= \$418.68 + \$1,008.88 \\
= \$1,427.56
\]

\[
200 \times \text{PacBio RS II C2} = \text{Library Prep} + 4 \text{ SMRT cells} \\
= \$418.68 + \$1,008.88 \\
= \$1,427.56
\]

\[
100 \times \text{PacBio RS II XL} = \text{Library Prep} \\
+ 2 \text{ SMRT cells} \\
= \$418.68 + \$504.44 \\
= \$923.12
\]

If yields improve further to 500 Mbp, one SMRT cell would become sufficient for 100× of XL-C2, at a cost of:

\[
100 \times \text{PacBio RS (proj.) XL-C2} = \text{Library Prep} + 1 \text{ SMRT cell} \\
= \$418.68 + \$252.22 = \$670.90
\]

For comparison, Riberio et al. [7] estimate the cost of closure at $13,124 assuming Illumina paired-end and jumping library preps and a resulting assembly with 50 gaps. Based on our simulation of NCBI genomes, if it is assumed all repeats below 500 bp are resolved, the average number of gaps per genome is 46 ± 52, with a maximum of over 500 gaps.

Data release

All data, supplementary files, assemblies, and code described here are available at [59].

The sequencing data and assemblies for novel strains generated for this study have been deposited in NCBI under the following accessions: *E. coli* K12 [PRJNA194437], *B. trehalosi* [PRJNA157929], *M. haemolytica* [PRJNA212438], *S. enterica* [PRJNA51643], *E. coli* O157:H7 [PRJNA63279], and *F. tularensis* [PRJNA212941]. The Illumina MiSeq *E. coli* K12 sequencing data are available from the Illumina Scientific Data Website [60].

Additional files

**Additional file 1: Repeat Stats.** Listing of repeat count and maximum repeat size for all complete genomes.

**Additional file 2: Report.log.krona.** An interactive Krona [35] chart detailing the expected number of gaps for all strains, species, genera, and so on under various coverage and chemistry scenarios.

**Additional file 3: ValidationResults.** Reference-free assembly validation. The columns are as described in Table 2 with the addition of the max contig, unmated LAP score, FRC feature count, and ALE scores.

Abbreviations

bp: base pair; CCS: Circular consensus sequencing; CLR: Continuous long read; indel: Insertion/deletion; PBcR: PacBio corrected read; QV: ‘Phred’s’ quality value; SNP: Single-nucleotide polymorphism.

Competing interests

The authors declared that they have no competing interests.

Authors’ contributions

SK and AMP conceived the project. SK implemented the algorithms and ran experiments. SK and AMP wrote the manuscript and generated figures. GPH performed manual validation of assemblies. DH, JB, DSM, NHB, and TS isolated strains and performed sequencing. All authors read and approved the final manuscript.

Acknowledgements

The authors thank Renee A Godtel and Robert T Lee for generating CLR, 454, and Illumina sequence data. We thank the NBACC Genomics group for generating CLR and 454 data. We thank Illumina Inc. for providing MiSeq sequencing data. We also thank Todd J Treangen and Brian Ondov for their careful reading of the draft and valuable discussions. We thank Pacific Biosciences for providing E. coli K12 sequencing data. Aaron Klammer, Jason Chin, David Alexander, and Edwin Hauw in particular for useful assistance.

Additional files

**Additional file 1: Repeat Stats.** Listing of repeat count and maximum repeat size for all complete genomes.

**Additional file 2: Report.log.krona.** An interactive Krona [35] chart detailing the expected number of gaps for all strains, species, genera, and so on under various coverage and chemistry scenarios.

**Additional file 3: ValidationResults.** Reference-free assembly validation. The columns are as described in Table 2 with the addition of the max contig, unmated LAP score, FRC feature count, and ALE scores.

Abbreviations

bp: base pair; CCS: Circular consensus sequencing; CLR: Continuous long read; indel: Insertion/deletion; PBcR: PacBio corrected read; QV: ‘Phred’s’ quality value; SNP: Single-nucleotide polymorphism.

Competing interests

The authors declared that they have no competing interests.

Authors’ contributions

SK and AMP conceived the project. SK implemented the algorithms and ran experiments. SK and AMP wrote the manuscript and generated figures. GPH performed manual validation of assemblies. DH, JB, DSM, NHB, and TS isolated strains and provided sequencing. All authors read and approved the final manuscript.

Acknowledgements

The authors thank Renee A Godtel and Robert T Lee for generating CLR, 454, and Illumina sequence data. We thank the NBACC Genomics group for generating CLR and 454 data. We thank Illumina Inc. for providing MiSeq sequencing data. We also thank Todd J Treangen and Brian Ondov for their careful reading of the draft and valuable discussions. We thank Pacific Biosciences for providing *E. coli* K12 sequencing data. Aaron Klammer, Jason Chin, David Alexander, and Edwin Hauw in particular for useful assistance.
and discussions. We thank David A Rasko for providing E. coli K12 MG1655 DNA. We thank Mick Watson for feedback on a pre-print version of the manuscript. The contributions of SK, DR, NH, and AMP were funded under Agreement No. H5QDCC-07-C-00020 awarded by the Department of Homeland Security (DHS) for the management and operation of the National Biodefense Analysis and Countermeasures Center (NBACC), a Federally Funded Research and Development Center. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the US Department of Homeland Security. In no event shall the DHS, NBACC, or Battelle National Biodefense Institute (BNBI) have any responsibility or liability for any use, misuse, inability to use, or reliance upon the information contained herein. The Department of Homeland Security does not endorse any products or commercial services mentioned in this publication. The contributions of GPH, JB, DSM, DH, TPLS were funded by United States Department of Agriculture, Agricultural Research Service. The contributions of DSM were funded by the Nebraska Agriculture Experiment Station (USDA Formula Funds and UNL funds) and the Nebraska Veterinary Diagnostic Laboratory. Genome sequencing of the E. coli O157:H7 and s. enteritidis Newport strains was supported in part by The Beef Checkoff (JLB and DBH). USDAs disclaimer: mention of a trade name, proprietary product, or specific equipment does not constitute a guarantee or warranty by the USDAs and does not imply approval to the exclusion of other products that may be suitable. The US Meat Animal Research Center is an equal opportunity employer and provider.

Author details
1 National Biodefense Analysis and Countermeasures Center, 110 Thomas Johnson Drive, Frederick, MD 21702, USA. 2 USDA, AMS, Meat Animal Research Center, Clay Center, NE 68933, USA. 3 USDA, ARS, Center for Grain and Animal Health Research, Manhattan, KS 66502, USA.

Received: 15 April 2013 Accepted: 13 September 2013 Published: 13 September 2013

References
1. Kyrpides NC: Genomes OnLine database (GOLD 1.0): A monitor of complete microbial genome sequencing (you get what you pay for). Nucleic Acids Res 2012, 40:D571–579.
2. Fraser CM, Eisen JA, Nelson KE, Paulsen IT, Salzberg SL: The value of complete microbial genome sequencing (you get what you pay for). J Bacteriol 2002, 184:6405–6408.
3. Pham CS, Grahn DJ, Fulton RS, Fitzgerald MG, Hostetter L, Muzny D, Ali I, Biren B, Bruce DC, Buhay C, Cole JR, Ding Y, Dugan S, Field D, Garrity GM, Gibbs R, Graves B, Han CS, Harrison SH, Highlander S, Hugenholtz P, Khouri HM, Kodina CD, Koller E, Kyrpides NC, Lang D, Lapidus A, Maitland SA, Markowitz VM, Mithra T, et al: Genomics. Genome project standards in a new era of sequencing. Science 2009, 326:236–237.
4. Salzberg SL, York JA: Beware of mis-assembled genomes. Bioinformatics 2005, 21:4320–4321.
5. Bashir A, Klammer AA, Robbins WP, Chin CS, Webster D, Paxinos E, Hsu D, Ashby M, Wang S, Peluso P, Sebra R, Sorenson J, Bullard J, Yen J, Valdivoino M, Mollavou E, Luong K, Lin S, Latmay B, Joshi A, Rowe L, Face M, Tart CL, Tumseklk M, Davis BM, Kasarskis A, Mekalans J, Ilevord MK, Schadt EE: A hybrid approach for the automated finishing of bacterial genomes. Nat Biotechnol 2013, 31:457–461.
6. Ribeiro FJ, Pribylsky D, Yin S, Sharpe T, Gerner S, Abouelell S, Berlin AM, Montmaurer A, Shep TA, Walker DR, Young SK, Russ C, Nusbaum C, MacCallum I, Jaffe DB: Finished bacterial genomes from single-molecule sequencing data. Genome Res 2012, 22:2270–2277.
7. Nagarajan N, Pop M: Sequence assembly demystified. Nat Rev Genet 2013, 14:517–567.
8. Hooper SD, Berg OG: On the nature of gene innovation: duplication and de novo assembly of single-molecule nanopore sequencing reads. Nat Nanotechnol 2009, 4:265–270.
9. Timp W, Mira©ov M, Wang D, Corner J, Aksamitiev A, Timp G: Nanopore sequencing: electrical measurements of the code of life. IEEE Trans Nanotechnol 2010, 9:281–294.
10. Branton D, Deamer DW, Marzali A, Bayley H, Bennner SA, Butler T, D'Ventra M, Garaj S, Higgs A, Huang X, Jovanovich SB, Kritzic PS, Lindsay S, Ling XS, Mastrangelo CH, Meller A, Oliver JS, Pershin YV, Ramsey JM, Riehn R, Soni GV, Tabard-Cossa V, Wanunu M, Wiggins M, Schloss JA: The potential and challenges of nanopore sequencing. Nat Biotechnol 2008, 26:146–1513.
11. Clarke J, Wu YC, Jhamlikghe L, Patel A, Reid S, Bayley H: Continuous identification of single-molecule nanopore DNA sequencing. Nat Nanotechnol 2009, 4:265–270.
12. Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, Peluso P, Rank D, Baybayan P, Bettmann B, Bibiolo A, Bjornson K, Chahudiri B, Christians F, Cicero R, Clark S, Dalal R, Dewinter A, Dixon J, Foquet M, Gaetner A, Hardenbol P, Heiner C, Hester K, Holden D, Kearns G, Kong X, Kuse R, Lacroix Y, Lin S, et al: Real-time DNA sequencing from single polymerase molecules. Science 2009, 323:133–138.
13. Chin CS, Sorenson J, Harris JB, Robbins WP, Charles RC, Jean-Charles RR, Bullard J, Webster DR, Kasarskis A, Peluso P, Paxinos EE, Yamauchi Y, Caldenwood SB, Mekalans JJ, Schadt EE, Waldor MK: The origin of the Haitian cholera outbreak strain. N Engl J Med 2011, 364:83–92.
14. Rasko DA, Webster DR, Sahi J, Bashir A, Boisen N, Schuetz F, Paxinos EE, Sebra R, Chin CS, Iliopoulos D, Klummer A, Peluso P, Lee L, Kistyluk AO, Bullard J, Kasarskis A, Wang S, Eid J, Rank D, Redman JC, Steyer R, Frimodt-Moller J, Cruse P, Petersen AM, Krogtell KG, Nataro JP, Schadt EE, Waldor MK: Origins of the E. coli strain causing an outbreak of hemolytic-uremic syndrome in Germany. N Engl J Med 2011, 365:709–717.
15. Fang G, Munera D, Friedman DJ, Mandlik A, Chao MC, Warerjee O, Feng Z, Loix B, Mahajan MC, Jlabdo DJ, Dekus G, Clark TA, Luong K, Murray IA, Davis BM, Bern-Paz A, Chess A, Roberts AJ, Krollach J, Turner SW, Kurnar V, Waldor MK, Schadt EE: Genome-wide mapping of methadined adenine residues in pathogenic Escherichia coli using single-molecule real-time sequencing. Nat Biotechnol 2013, 31:1232–1239.
16. English AC, Richards S, Han Y, Wang M, Vee V, Guo J, Qin X, Muzny DM, Reid JG, Worley KC, Gibbs RA: Mind the gap: upgrading genomes with Pacific Bioscences RS long-read sequencing technology. PLoS One 2013, 7:e47768.
17. Ono Y, Arai K, Hamada M: PSBIM: PacBio reads simulator–toward accurate genome assembly. Bioinformatics 2013, 29:119–121.
18. Waldbieser G, Kertesz M, Puslakre D, Bawinkamp T, Liu J: Production of long (1.5 kb–150 kb) accurate, DNA sequencing reads using an illumina HSq2000 to support de novo assembly of the blue catfish genome. San Diego, CA: Plant and Animal Genome XXII Conference; 2013. https:// proteomics.ca/pag/xxi/webprogram/Paper7088.html.
19. Miller JR, Koren S, Sutton G: Assembly algorithms for next-generation sequencing data. Genomics 2010, 95:315–327.
20. Pop M: Genome assembly reborn: recent computational challenges. Brief Bioinformatics 2009, 10:354–366.
21. Nagarajan N, Pop M: Sequence assembly demystified. Nat Rev Genet 2013, 14:517–567.
22. Phillippy AM, Schatz MC, Pop M: Genome assembly forensics: finding the elusive mis-assembly. Genome Biol 2008, 9:R55–R65.
23. Hooper SD, Berg OG: On the nature of gene innovation: duplication patterns in microbial genomes. Mol Biol Evol 2003, 20:945–954.
24. Lerat E, Daubin V, Ochman H, Moran NA: Evolutionary origins of genomic repertoires in bacteria. PLoS Biol 2005, 3:e130.
25. Barrangou R, Fremaux C, Deveau H, Richards M, Boyaval P, Moinae S, Romero DA, Horvath P: CRISPR provides acquired resistance against viruses in prokaryotes. Science 2007, 315:1709–1712.
26. Treangen TJ, Abraham A-L, Toucaon M, Rocha EP: Genesis, effects and fates of repeats in prokayotic genes. FEMS Microbiol Rev 2009, 33:559–571.
27. Kingsford C, Schatz MC, Pop M: Assembly complexity of prokaryotic genomes using short reads. BMC Bioinformatics 2010, 11:21.
28. Chaisson MJ, Tesker G: Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. BMC Bioinformatics 2012, 13:238.
29. Lee H, Schatz MC: Genomic dark matter: the reliability of short read mapping illustrated by the genome mappability score. Bioinformatics 2013, 29:2097–2103.
30. Pacific Biosciences Delivers Enhanced DNA Sequencing Chemistry and Software to Help Solve Complex Genetic Problems. http://investor.pacificbiosciences.com/releasetext.cfm?ReleaseId=718855.
31. Miller JR, Delcher AL, Koren S, Venter E, Walenz BP, Brownley A, Johnson J, Li K, Mobarry C, Sutton G: Aggressive assembly of pyrosequencing reads with mates. Bioinformatics 2008, 24:2818–2824.
