Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data

Article *in* Nature Methods · May 2013

DOI: 10.1038/nmeth.2474 · Source: PubMed

12 authors, including:

Aaron Klammer
Pacific Biosciences of California, Inc.
13 PUBLICATIONS **1,451** CITATIONS
SEE PROFILE

Cheryl Heiner
Pacific Biosciences of California, Inc.
29 PUBLICATIONS **17,105** CITATIONS
SEE PROFILE

Alex C Copeland
Lawrence Berkeley National Laboratory
305 PUBLICATIONS **7,156** CITATIONS
SEE PROFILE

John Lawton Huddleston
University of Washington Seattle
75 PUBLICATIONS **1,036** CITATIONS
SEE PROFILE

Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data

Chen-Shan Chin[1], David H Alexander[1], Patrick Marks[1], Aaron A Klammer[1], James Drake[1], Cheryl Heiner[1], Alicia Clum[2], Alex Copeland[2], John Huddleston[3], Evan E Eichler[3], Stephen W Turner[1] & Jonas Korlach[1]

We present a hierarchical genome-assembly process (HGAP) for high-quality *de novo* microbial genome assemblies using only a single, long-insert shotgun DNA library in conjunction with Single Molecule, Real-Time (SMRT) DNA sequencing. Our method uses the longest reads as seeds to recruit all other reads for construction of highly accurate preassembled reads through a directed acyclic graph–based consensus procedure, which we follow with assembly using off-the-shelf long-read assemblers. In contrast to hybrid approaches, HGAP does not require highly accurate raw reads for error correction. We demonstrate efficient genome assembly for several microorganisms using as few as three SMRT Cell zero-mode waveguide arrays of sequencing and for BACs using just one SMRT Cell. Long repeat regions can be successfully resolved with this workflow. We also describe a consensus algorithm that incorporates SMRT sequencing primary quality values to produce *de novo* genome sequence exceeding 99.999% accuracy.

Determining the genomic sequences of microorganisms is a prerequisite to understanding their biology, rationally manipulating their function, tracing their history and geographical distribution, and—in the case of pathogens—developing effective treatments[1–3]. Sequence information can be obtained rapidly and cost-effectively using second-generation sequencing techniques, and differences between isolates can be detected by comparing sequence reads to related reference strains[4]. However, this information provides an incomplete view into the genomes of microorganisms under study because it is restricted to what is known for the reference strain that is used for comparison. Certain aspects of microbial diversity, including large-scale structural rearrangements, segmental duplications or inversions, and horizontal transfer of mobile elements such as phages and plasmids, can be overlooked by these types of resequencing approaches because the sequence reads that do not align to the chosen reference genome are removed from consideration. In many cases, it is those sequences that provide critical insights into what makes certain bacterial strains different from their reference strains[5–8].

Obtaining the complete genome sequence of microbes in an automated, high-throughput manner has been challenging[5,9,10]. Because of the short read lengths in second-generation sequencing methods, long repeats present in multiple copies often cannot be resolved, resulting in unfinished, fragmented draft assemblies[11]. Gaps in draft genome assemblies can also be caused by extreme sequence contexts such as GC- or AT-rich regions or palindromic sequences, both of which are frequently not covered by second-generation sequencing methods. Because of these limitations, Sanger sequencing has typically been used to 'finish' microbial genomes, but the laborious and low-throughput nature of this process make it slow and expensive. Therefore, efficient methods for finished, high-quality *de novo* genome determinations that do not rely on assumptions about the DNA sample under study are highly desirable for capturing the complete genetic constitution of microorganisms in an unbiased, hypothesis-free manner.

Recently, SMRT DNA sequencing (Pacific Biosciences) has been used to generate sequencing reads that are much longer than second-generation or even Sanger sequencing reads, facilitating *de novo* genome assembly and genome finishing[12,13]. For typical bacterial genome sizes (1–10 megabases (Mb)), hybrid assembly approaches that use the long SMRT sequencing reads in conjunction with shorter reads (from SMRT circular consensus sequencing reads or second-generation sequencing methods) have been used to generate finished, high-quality genome assemblies in automated workflows[14–16]. In these hybrid strategies, the short reads are used to correct errors in the long SMRT sequencing reads; the corrected long reads are then subjected to a kilobase-read, overlap-based assembler. Although these strategies have been applied successfully to a variety of microbes and eukaryotic organisms, hybrid assembly requires the preparation of at least two different sequencing libraries and several types of sequencing runs (and sometimes multiple sequencing platforms). For more efficient genome closing, a homogeneous workflow requiring only one library and sequencing method is desirable.

Hybrid assemblies are known to break in regions with insufficient second-generation sequencing data coverage (owing to the GC or sequence context biases mentioned above). For those regions, it was found that a consensus established from

[1]Pacific Biosciences, Menlo Park, California, USA. [2]Joint Genome Institute, Walnut Creek, California, USA. [3]Department of Genome Sciences, University of Washington, Seattle, Washington, USA. Correspondence should be addressed to J.K. (jkorlach@pacificbiosciences.com).

RECEIVED 31 JANUARY; ACCEPTED 4 APRIL; PUBLISHED ONLINE 5 MAY 2013; DOI:10.1038/NMETH.2474

the SMRT sequencing reads mapping to these uncovered regions could be used to span them and thereby connect more contigs[15]—a process that can be implemented in the latest version (7.0) of the Celera Assembler.

Extending this principle genome wide, we reasoned that we could leverage the long reads, lack of bias and high consensus accuracy due to the random nature of errors in SMRT sequencing to generate finished genomes using long insert–library SMRT sequencing exclusively. To achieve this, we developed a consensus algorithm that preassembles long and highly accurate overlapping sequences by correcting errors on the longest reads using shorter reads from the same library. We describe a nonhybrid HGAP that implements this approach in a fully automated workflow, and we demonstrate the *de novo* construction of several microbial genomes into finished, single-contig assemblies. We evaluate the performance of the method on several bacterial genomes for which Sanger and 454 (Roche) sequencing were used to generate reference sequences, finding that the *de novo* assembly is collinear with these references and >99.999% (quality value (QV) of >50) concordant. We also show that HGAP can be used to sequence and effectively assemble BACs containing sequences that can be problematic for second-generation approaches.

RESULTS

Hierarchical genome-assembly process workflow

The principle (**Fig. 1**) and workflow (**Fig. 2**) of HGAP consist of several well-defined steps. (1) Select the longest sequencing reads as a seeding sequence data set. (2) Use each seeding sequence as a reference to recruit shorter reads, and preassemble reads through a consensus procedure. (3) Assemble the preassembled reads using an off-the-shelf assembler that can accept long reads. (4) Refine the assembly using all initial read data to generate the final consensus that represents the genome. Optionally, minimus2 or similar tools can be used to connect the contigs from step (3) to further improve the continuity of the assembly and remove spurious contigs due to assembly or sequencing errors[17,18].

The preassembly step converts long raw SMRT sequencing reads into high-quality sequences that can be used with existing long-read assemblers. It is based on alignment and mapping between raw reads in combination with a directed acyclic graph–based consensus step to remove randomly distributed deletion and insertion errors. It is conceptually akin to a mini-assembly process to generate highly accurate preassembled reads. In addition, low-quality and chimeric sequence reads are removed during this process. In contrast to hybrid approaches, HGAP does not require highly accurate raw reads for error correction.

Details of each analysis step are described in the Online Methods. The technical algorithm and implementation details are provided in **Supplementary Note 1**.

Application to *Escherichia coli*

To evaluate HGAP, we first applied it to *E. coli* K-12 MG1655, for which a high-quality reference sequence had previously been generated by Sanger sequencing (NC_000913.2, genome size 4,639,675 base pairs (bp))[19]. We prepared and sequenced a single ~8.5-kilobase (kb) SMRTbell library (**Fig. 2**) using eight SMRT Cells, which yielded 461 Mb of sequence from 141,492 continuous long reads, with a typical average read length of 3,257 bp. The data

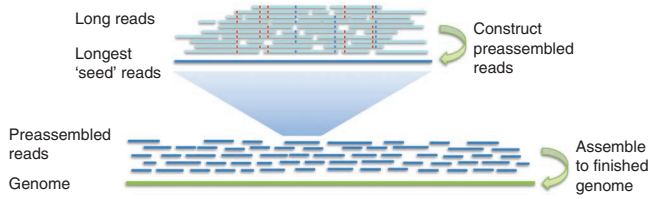

Figure 1 | Principle of the hierarchical genome-assembly process using long-insert-size DNA shotgun template libraries with SMRT sequencing. The longest reads are selected as 'seed' reads, to which all other reads are mapped. A preassembly is performed that converts the seed reads into highly accurate preassembled reads that are used for the genome assembly, which is followed by a final consensus-calling step (not shown).

were then subjected to the HGAP method (the optional cleanup by minimus2 or similar tools was not used in this study to show the unaltered output of HGAP).

The availability of a high-quality reference allowed us to characterize the algorithm at each assembly step. First, we examined the length and accuracy of the seed reads by aligning each read longer than the 6-kb cutoff to the reference sequence (**Fig. 2** and **Supplementary Fig. 1a**). We found that 17,726 seed reads representing ~140 Mb of total sequence fulfilled this criterion and had an average aligned read length and single-pass accuracy of 7,213 bp and 86.9%, respectively. The aligned read length was shorter than the overall mean seed read length (8,160 bp) because, for some reads, lower-quality regions or chimeric reads did not align.

In the preassembly stage (see Online Methods), the seed reads were converted into 17,232 highly accurate preassembled reads with a mean length of 5,777 bp and a mean accuracy of 99.9% (**Fig. 2** and **Supplementary Fig. 1b**). The drop in read length is due to end trimming and filtering of spurious and chimeric reads. About 30%–35% of total bases are typically removed during preassembly as a function of the mapping and trimming parameters, but these parameters can be further optimized to improve the yield and length distribution in future implementations of HGAP.

Subjecting the preassembled reads to the Celera Assembler yielded one 4,656,144-bp contig representing the *E. coli* genome and a spurious small 7,589-bp contig (aligning to positions 2,393,788–2,401,380 of the reference with 99.96% identity). A genome-wide alignment showed that the assembly spanned the entire *E. coli* reference and was collinear with it (**Fig. 2**; see **Table 1** for final assembly statistics and **Supplementary Table 1** for detailed statistics). The Celera Assembler assumes the genome to be linear, so the assembly was slightly larger (100.35%) than the reference because the ends of the single contig that cover a genome can have overlaps. These overlaps can be trimmed manually on the basis of the sequence identity of the overlapping end regions (data not shown) to give a finished genome sequence that is the same size as the reference sequence.

The consensus accuracy of the assembly was evaluated with genome-wide alignments using the MUMmer package[20] (**Supplementary Note 1**). Because not all potential sequencing errors are removed in the preassembly step, it is likely that the output from the Celera Assembler still contains some errors. To evaluate the effect of the Quiver consensus algorithm on the final quality of the assembly, we compared the reference concordance before and after the Quiver consensus step (**Supplementary Table 1**).

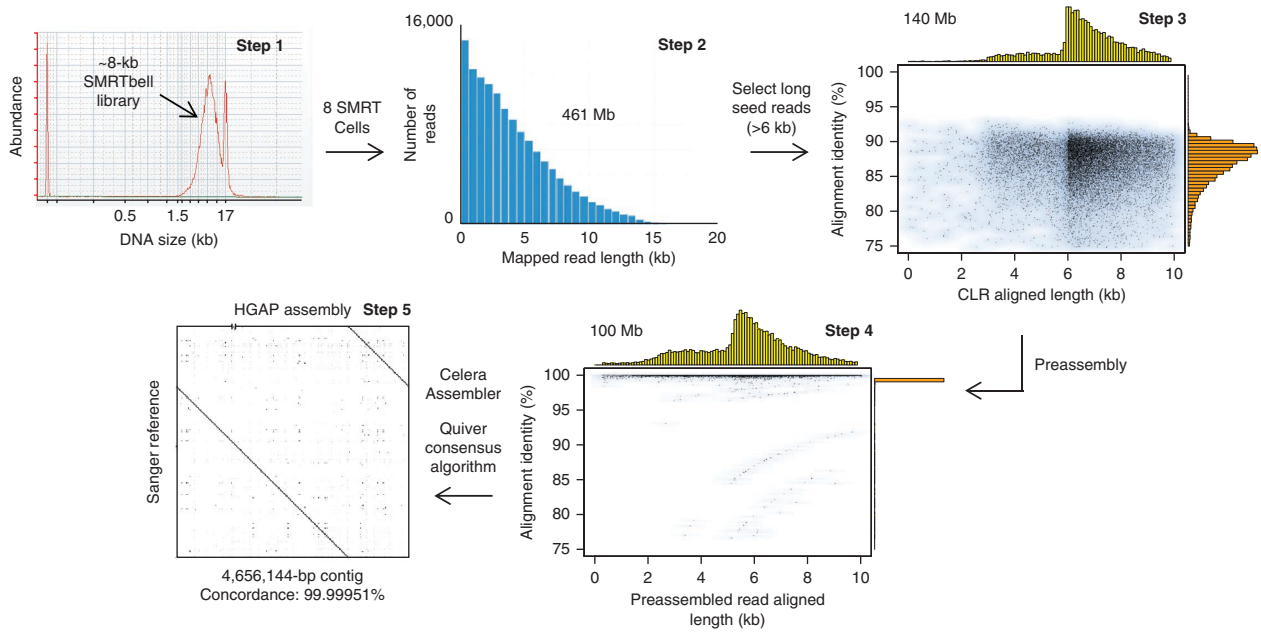

Figure 2 | Workflow for the *de novo* HGAP assembly of *E. coli* MG1655. Data for steps 3 and 4 are provided for all microorganisms studied in **Supplementary Figures 1–3**. CLR, continuous long read.

Quiver reduced the number of differences with the Sanger-based reference from 49 in the initial assembly to 23, corresponding to a nominal QV of 53 (99.9995% identity).

Of the remaining 23 differences, 9 point mutations were validated with PCR-based Sanger sequencing as biological variations in the *E. coli* sample used in this study (**Supplementary Table 2**).

Table 1 | HGAP assembly statistics summary for three different microorganisms and one human BAC

SMRT Cells	CLR bases (Mb)	Assembly size (bp)	Number of contigs >10 kb; (total)	Assembly size relative to reference (%)	N50	Concordance with Sanger reference (%)	Nominal QV	Genes predicted (%)	Assembler
Escherichia coli MG1655									
8	461	4,656,144	1 (2)	100.35	4,648,564	99.99951	53.1	99.3	Celera
8	461	4,784,874	8 (16)	103.13	4,606,235	99.99937	52.0	99.1	MIRA
6	341	4,701,623	10 (14)	101.34	1,163,944	99.99938	52.1	99.0	Celera
6	341	5,043,988	26 (52)	108.71	455,003	99.99939	52.1	98.6	MIRA
4	232	4,689,701	17 (21)	101.08	392,114	99.99876	49.1	98.2	Celera
4	232	4,807,190	25 (42)	103.61	317,682	99.99906	50.3	97.7	MIRA
Meiothermus ruber DSM1279									
4	334	3,098,781	1	100.04	3,098,781	99.99965	54.5	99.3	Celera
4	334	3,134,158	1 (5)	101.18	3,103,747	99.99978	56.5	99.5	MIRA
3	248	3,098,729	1	100.04	3,098,729	99.99958	53.8	99.2	Celera
3	248	3,154,602	4 (7)	101.84	3,101,561	99.99968	55.0	99.3	MIRA
2	170	3,102,769	3	100.17	1,053,479	99.99897	49.9	98.8	Celera
2	170	3,138,573	4 (5)	101.33	3,096,314	99.99939	52.2	99.0	MIRA
Pedobacter heparinus DSM2366									
7	485	5,171,533	2 (3)	100.08	2,927,691	99.99959	53.9	99.4	Celera
7	485	5,197,624	1 (5)	100.59	5,164,849	99.99960	53.9	99.3	MIRA
6	408	5,173,388	2 (3)	100.12	2,928,902	99.99969	55.1	99.3	Celera
6	408	5,174,349	2 (3)	100.13	3,511,353	99.99969	55.1	99.3	MIRA
4	274	5,184,825	11 (18)	100.34	1,403,814	99.99944	52.5	98.9	Celera
4	274	5,196,690	15 (22)	100.57	1,258,275	99.99950	53.0	98.6	MIRA
Human BAC (VMRC53-364D19)									
1	85	186,053	1 (4[a])	100.00	186,053	N/A	N/A	N/A	Celera

For full statistics, see **Supplementary Table 1**. CLR, continuous long read; N50, *N* such that 50% of the bases in the assembly are contained in contigs ≥ *N* ; QV, quality value.
[a]The three additional contigs were the result of *E. coli* contamination.

Table 2 | Comparison of the *E. coli* HGAP assembly of this study to earlier hybrid assembly approaches

Study	Method	Illumina library and data details	PacBio library and data details	Assembly size (bp)	Number of contigs	N50	Reported base concordance (%)
Ref. 16	ALLPATHS-LG	239,610,582-bp (2,372,382 reads), 180-bp-insert paired-end library; 367,889,95-bp (3,955,806 reads), ~3-kb jumping library	C1 chemistry 619,784,574 bp (409,304 reads) Median length = 1,261 bp Maximum length = 9,724 bp	4,638,970	1	4,638,970	99.999957 (2 errors)
Ref. 15	PacBioToCA with Celera Assembler	22,720,100 reads of 100 bp, 500-bp-insert paired-end library	Data collected with preleased instrument 251,762 reads Median length = 540 bp Maximum length = 3,787 bp	4,465,533	77	89,431	99.99916 (39 differences)
This study (eight SMRT Cells)	HGAP with Celera Assembler	–	10-kb SMRTbell insert, XL/C2 chemistry 460,967,046 bp (141,492 reads) Median length = 2,755 bp Maximum length = 17,831 bp	4,656,144	2	4,648,564	99.99951 (23 differences, 14 errors)

Ribeiro *et al.*[16] used long Pacific Biosciences (PacBio) reads to resolve midrange ambiguities and to fill gaps in an initial short-read assembly that was constructed using a modified de Bruijn graph approach. The PacBio library was constructed with shorter inserts and sequenced with an earlier chemistry, and longer-range information was derived from an ~3-kb jumping Illumina library. Koren *et al.*[15] used PacBioToCA to correct PacBio reads before assembling with the Celera Assembler. No final consensus was generated using PacBio data, and reads were substantially shorter than those from the current study as data were collected using a prerelease instrument and sequencing chemistry. The reference genome size is 4,639,675 bp.

In addition, we observed five local structure variations. SMRT sequence reads mapping to all five structural-variation regions showed that the long-read data were consistent with the structural variations inferred from our assembly, indicating that this particular strain of *E. coli* K-12 MG1655 differed from the reference strain in these regions (**Supplementary Figs. 4–8**). The high accuracy of the final assembly allowed the correct prediction of over 99% of all genes present in the organism (**Table 1**).

Assembly quality

We compared the HGAP assembly to previously described hybrid technology–based assembly approaches (**Table 2**). The assembly described here is of the same contiguity as, and has similar accuracy to, an approach combining long SMRT sequencing reads with two additional Illumina libraries (a 0.2-kb library and an ~3-kb jumping library) subjected to Illumina sequencing[16].

We also explored assembly quality as a function of different amounts of sequencing data (**Table 1**, **Supplementary Table 1** and **Supplementary Fig. 9**). With only six and then four SMRT Cells worth of data, HGAP returned 7 and 17 contigs above 0.1 Mb (14 and 21 total), respectively. In all cases, >98.9% of all genes were correctly predicted from the assemblies. Therefore, for certain applications in which close-to-finished genomes are sufficient, it may be more economical to run a smaller number of SMRT Cells.

Resolving repeat regions

One of the most important prerequisites for obtaining high-quality, finished genome assemblies is the resolution of repeat regions. Long, exact repeats can cause misassemblies or fragmented assemblies if the sequence reads are not long enough to span the repeats with unique flanking sequences. For non-exact repeats, it is important that sequence reads are correctly discerned from different instances of a repeat. This is important not only for the assembler to generate correct contigs but also for the Quiver consensus algorithm to produce the most accurate sequence: if unresolved or collapsed repeats exist in the final genome assembly, the consensus algorithm, which relies on unambiguous mapping, might not generate optimum results because of collapsed repeats.

We examined HGAP's capability to resolve repeats by comparing the highly similar rRNA operon repeats in *E. coli* MG1655. We selected two seed reads that correspond to repeat regions within positions 2,724,000–2,734,436 and 3,420,825–3,431,015 of the reference and have 95.9% identity over 5,404 bp. Four types of reads map to each seed read (**Supplementary Fig. 10**): (i) reads that are anchored by unique regions at both ends, (ii) reads that are anchored at one end only, (iii) reads that map fully within the repeat region and (iv) reads that do not overlap with the repeat regions.

Type (i) and (ii) reads are unlikely to be mismapped because they have unique flanking sequences. When a repeat is smaller than the average read length, most of the reads will be types (i) and (ii), generating excellent assemblies and consensus accuracies regardless of the similarity to other genomic copies of the repeat. In the case of long repeat regions, more reads will be of type (iii), for which only real biological differences between repeat copies can be used to avoid mismapping. We found that because sequencing errors were random and did not correlate with the differences between repeats, most type (iii) reads still mapped correctly. To demonstrate this, we aligned all reads from the two repeat regions to the Sanger reference or seed reads (**Supplementary Fig. 11**) and compared alignment scores. Although the two seed reads had accuracies of 88.4% and 89.8%, which are below the 95.9% identity between repeats, the read alignment scores of type (iii) reads had sufficient sensitivity to distinguish them and map correctly.

To determine whether errors are more likely to be encountered in repetitive regions, we examined assembly quality across the rRNA repeats for the *E. coli* sample. We examined the 22 annotated rRNA repeats with flanking regions (500 bp on both ends). The regions ranged from 1.1 kb to 3.9 kb, and some occurred in tandem, with similarity from 99.4% (9 differences out of 1,540 bp) to 100%. Out of the 54 kb of total sequence, there were only two differences between the assembly and the reference, meaning that

the repeats were resolved correctly. The assemblies characterized by many reads anchored at unique regions contained few errors due to paralogous or repetitive regions.

In SMRT sequencing, more data can be collected to increase the chances of obtaining reads that span through the longest repeats to resolve them. As an example, most contigs in the four–SMRT Cell assembly described above (a total of 17 contigs over 10 kb; 21 contigs total) ended at the rRNA operon repeats. As expected, all rRNA operon repeats (~5 kb each) were successfully resolved in the two-contig eight–SMRT Cell assembly because the added sequencing data provided a greater fraction of long reads that cover these regions.

Application to other bacteria

We applied HGAP to two additional microorganisms, both of which had previously been sequenced (**Table 1**).

Meiothermus ruber DSM1279 is a nonmotile, aerobic, thermophilic bacterium isolated from a hot spring (Kamchatka peninsula, Russia) with a single replicon and a genome of 3,097,457 bp with a GC content of 63% and a repeat content of 1.7% (window size of 25 bases)[21]. We performed SMRT sequencing of an ~10-kb insert library of genomic DNA from this organism using four SMRT Cells and found that three SMRT Cells were sufficient to produce single-contig, finished genomes as uncurated HGAP output (**Table 1**). The assembly from four SMRT Cells contained one misassembly with respect to the reference, caused by a 7,092-bp exact repeat (**Supplementary Fig. 12**). Notably, when we applied the same method to data from three SMRT Cells, the single contig generated by the Celera Assembler agreed with the reference, but the start and end of the contig were broken at the second repeat copy.

Apart from the single misassembled repeat, the assemblies covered the entire *M. ruber* genome at ≥99.999% concordance with the Sanger-based reference (nominal QV >50), and over 99% of genes were predicted correctly (**Table 1**). Quiver improved accuracy markedly, correcting 130 bases in the initial assembly, predominantly in high-GC homonucleotide stretches. Of the residual 11 base differences from the reference, we chose 8 for validation by PCR and Sanger sequencing and found that 7 supported the SMRT sequencing call, which was indicative of biological variation relative to the reference rather than of sequencing errors (**Supplementary Table 3**).

Pedobacter heparinus DSM2366 is a motile, aerobic, mesophilic bacterium with a single replicon and a genome of 5,167,383 bases with a GC content of 42% and a repeat content of 1.5% (window size of 25 bases)[22]. It is the type species of the rapidly growing *Pedobacter* genus within the phylum Bacteroidetes. *P. heparinus* was the first isolated strain shown to grow with heparin as the sole carbon and nitrogen source, and it produces several enzymes involved in the degradation of mucopolysaccharides. We generated seven SMRT Cells of sequencing data and used four, six or all seven SMRT Cells' worth of data in the HGAP workflow. For six and seven SMRT Cells, we obtained two-contig assemblies that contained the entire genome and were over 99.999% concordant with the reference (**Table 1**, **Supplementary Fig. 13** and **Supplementary Table 4**). Only one 6,139-bp-long repeat was not sufficiently resolved in the assembly. Similarly to with *E. coli* and *M. ruber*, we successfully predicted >99% of the genes from the *P. heparinus* assemblies. For this organism, additional Illumina sequence data were available, which allowed a comparison of the HGAP assembly with previously described hybrid assembly approaches (**Supplementary Note 2**). Our HGAP assembly was generally more contiguous, but slightly less accurate, than some of the hybrid assembly results.

Using alternative assemblers

To demonstrate the generality of applying preassembled reads to a variety of long read–capable assemblers, we compared the results of using the Celera and MIRA assemblers (**Table 1** and **Supplementary Table 1**). As expected from the similar assembler designs, results were alike, in terms of both genome contiguity and sequence accuracy. For the seven–SMRT Cell MIRA case for *P. heparinus*, a single contig spanning the entire reference was obtained; however, the 6,139-bp repeat was not confidently resolved in this assembly (**Supplementary Fig. 14**), analogously to the three–SMRT Cell assembly for *M. ruber* described above with the Celera Assembler. We hypothesize that the greedy nature of the heuristic algorithm optimized for speed in the layout stage of both Celera Assembler and MIRA might not detect the inconsistent overlaps around the repeats to break contigs accordingly, especially in cases in which the sequence read lengths are on the same order as the lengths of the long repeats. It is possible to investigate the underlying overlap information of the DNA fragments used by an assembler so that such misassembled contigs can be broken properly. The probability of such misassemblies will be reduced with longer read lengths, provided that reads spanning unique parts on both ends of a repeat region are generated.

De novo BAC assembly

We also explored the utility of HGAP for other *de novo* assembly applications, including BACs, which are typically several hundred kilobases in size. Sanger sequencing of BAC libraries is the preferred approach to resolve genomic regions with high repeat content and structural variation in complex genomes, including the human genome[23–25]. For this demonstration, we chose an ~175-kb BAC (VMRC53-364D19) corresponding to a region on chromosome 15 (32,291,106–32,463,964 relative to the hg19 human reference sequence). Using a single SMRT Cell of sequencing, HGAP produced a correct and accurate BAC assembly (**Table 1**). From the initial 84.5 Mb of sequence, we chose reads longer than 3,200 bases as the seed reads (corresponding to 11.8 Mb of total sequence). After preassembly, we obtained 1,892 preassembled reads (5,027,597 bases total) with an average length of 2,657 bp. After assembly, we obtained four contigs of lengths 186,053 bp, 2,690 bp, 2,343 bp and 1,263 bp. The largest contig could be circularized and contained the targeted region and the BAC vector sequence; the smaller contigs were found to be small fragments of *E. coli* DNA sequences. In the final assembly, we observed 165 single-nucleotide polymorphisms relative to the hg19 reference sequence. All of the six single-nucleotide polymorphisms that we selected for PCR and Sanger sequencing were validated as true biological variations (**Supplementary Table 5**).

DISCUSSION

We have applied the *de novo* genome assembly method HGAP to three microbial genomes and one human BAC. In each case, SMRT sequencing of a single, large-insert template library to

~80×–100× coverage followed by HGAP analysis resulted in excellent *de novo* genome assemblies that are comparable in quality and contiguity to assemblies generated by Sanger sequencing or by hybrid approaches combining long- and short-read sequencing methods[16]. The assembly algorithm developed here simplifies the workflow from a microbial DNA sample to the finished genome considerably. Previously described hybrid technology–based algorithms required at least two separate sequencing libraries and sequencing runs, and in certain implementations, at least two different sequencing technologies[14–16]. Here, only a single, long-insert shotgun DNA library is prepared and subjected to automated continuous long-read SMRT sequencing, and the assembly is performed without the need for circular consensus sequencing[26].

Whereas the shorter reads from long-insert SMRT sequencing are discarded in the hybrid assembly approaches, the HGAP method described here uses every bit of such sequencing data: the longest reads for assembling a contiguous genome, and the shorter reads for improving the accuracy of the long reads through the preassembly process. Spurious, low-quality and chimeric reads are removed as early as possible in the overall genome assembly process and therefore do not contribute to artifactual contigs or misassemblies. The final consensus–calling algorithm Quiver, which takes into account all of the underlying data and the raw quality values inherent to SMRT sequencing, then polishes the assembly for final consensus accuracies in excess of 99.999% (QV of >50), which are on par with consensus accuracies achieved with Sanger sequencing[27,28]. The algorithm disregards the original assembly sequence for consensus calling and thus avoids artifacts caused by reference bias. It is worth noting that Quiver can also be used for resequencing applications (that is, when reads are mapped to a previously determined reference to identify variants), permitting variant detection with high sensitivity and specificity.

One prerequisite for obtaining finished, highly accurate genomes is the ability of a sequencing technology to generate high-quality sequence data over the entire range of sequence complexities and GC content present in that organism's genome. Systematic sequencing bias can introduce errors in the final genome sequence that cannot be overcome by additional sequencing coverage. SMRT sequencing has demonstrated excellent uniformity of sequence data over the widest range of GC content, including high-GC extremes at which other technologies, including even Sanger sequencing, fail[29]. Unbiased sequencing coverage allows for straightforward identification of any potential misassemblies or discordances with existing references, which appear as breaks in the random, shotgun-read structure (**Supplementary Figs. 4–8**). In addition, because sample preparation in SMRT sequencing does not include denaturation or amplification steps, palindromic sequences can be read successfully with this method, whereas they typically cause so-called 'hard stops' due to hairpin formation that are prohibitive in other sequencing systems[30].

Although single-pass accuracies in SMRT sequencing are lower than those of other, multimolecule-based sequencing technologies, the quality of the final genome sequence is determined in consensus. The random-error profile in SMRT sequencing and lack of systematic bias result in high consensus accuracies, as random errors wash out exponentially rapidly when consensus is generated[15]. In addition, the multikilobase read lengths in SMRT sequencing greatly facilitate the correct mapping of reads into repetitive regions by spanning the entire repeat units and anchoring the reads with unique flanking sequences. Difficulty around correctly mapping into repeat regions in genomes has been recognized as a source of errors in short-read sequencing technologies[14,31].

For smaller assembly targets such as BACs, given the current overall throughput of >100 Mb per SMRT Cell, it is likely that multiplex strategies to sequence pooled BACs could further enhance the efficiency of assemblies. Upon careful design of the sequencing run, one can potentially assemble all sequences at once if no long repeats between the different BACs in the pool are present and proper assembly preprocesses—for example, vector and *E. coli* contamination filtering—are included. Alternatively, barcoding strategies during library preparation could be used.

Our approach is currently capable of producing near-finished, high-quality genomes in terms of contiguity and error rate, and it should thereby facilitate closing of the large gap that currently exists between drafted and finished genomes[32]. The workflow described here is implemented in a fully automated process from DNA sample preparation to the determination of the finished genome and, in our hands, has been completed in as few as 4 d. In addition, the same SMRT sequencing reads can be used to determine the epigenome of the organism under study[33]. It will also be interesting to evaluate the utility of HGAP for the *de novo* assembly of eukaryotic genomes.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. The raw data for this study are deposited in NCBI Sequence Read Archive under BioProject numbers PRJNA196342 (*E. coli*; SRR811719, SRR811720, SRR811743–SRR811747 and SRR811770), PRJNA196343 (*M. ruber*; SRR811863–SRR811865 and SRR811890) and PRJNA196344 (*P. heparinus*; SRR811935–SRR811937, SRR811960–RR811963, SRR812176 and SRR812197).

Note: Supplementary information is available in the online version of the paper.

ACKNOWLEDGMENTS

We thank S. Clingenpeel (Joint Genome Institute) for growing cultures and performing DNA extraction for *M. ruber* and *P. heparinus*; B. Munson and F. Antonacci for assistance with the BAC library construction; and K. Travers, S. McCalmon, M. Wang, U. Nguyen, S. Ranade, M. Ashby, L. Hon and L. Hickey (Pacific Biosciences) for assistance in sample preparation, sequencing and data analysis. The authors acknowledge the ATCC for providing the *E. coli* K-12 MG1655 strain. We thank S. Koren and A. Phillippy for pointing out to us the SMRT sequencing–based gap-filling functionality development in the Celera Assembler. The work conducted by the US Department of Energy Joint Genome Institute is supported by the Office of Science of the US Department of Energy under contract no. DE-AC02-05CH11231.

AUTHOR CONTRIBUTIONS

C.-S.C., A. Copeland., E.E.E., S.W.T. & J.K. designed the experiments; C.-S.C., D.H.A., P.M., A.A.K., J.D., A. Clum and J.H. analyzed data; C.H. performed the validation sequencing; and C.-S.C., D.H.A., P.M., A.A.K., A. Copeland., E.E.E. and J.K. wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details are available in the online version of the paper.

1. Medini, D. *et al.* Microbiology in the post-genomic era. *Nat. Rev. Microbiol.* **6**, 419–430 (2008).
2. Parkhill, J. & Wren, B.W. Bacterial epidemiology and biology—lessons from genome sequencing. *Genome Biol.* **12**, 230 (2011).
3. Gagarinova, A. & Emili, A. Genome-scale genetic manipulation methods for exploring bacterial molecular biology. *Mol. Biosyst.* **8**, 1626–1638 (2012).
4. Loman, N.J. *et al.* High-throughput bacterial genome sequencing: an embarrassment of choice, a world of opportunity. *Nat. Rev. Microbiol.* **10**, 599–606 (2012).
5. Ricker, N., Qian, H. & Fulthorpe, R.R. The limitations of draft assemblies for understanding prokaryotic adaptation and evolution. *Genomics* **100**, 167–175 (2012).
6. Siguier, P., Filée, J. & Chandler, M. Insertion sequences in prokaryotic genomes. *Curr. Opin. Microbiol.* **9**, 526–531 (2006).
7. Srikhanta, Y.N., Fox, K.L. & Jennings, M.P. The phasevarion: phase variation of type III DNA methyltransferases controls coordinated switching in multiple genes. *Nat. Rev. Microbiol.* **8**, 196–206 (2010).
8. Toussaint, A. & Chandler, M. Prokaryote genome fluidity: toward a system approach of the mobilome. *Methods Mol. Biol.* **804**, 57–80 (2012).
9. Kingsford, C., Schatz, M.C. & Pop, M. Assembly complexity of prokaryotic genomes using short reads. *BMC Bioinformatics* **11**, 21 (2010).
10. Salzberg, S.L. *et al.* GAGE: a critical evaluation of genome assemblies and assembly algorithms. *Genome Res.* **22**, 557–567 (2012).
11. Fraser, C.M., Eisen, J.A., Nelson, K.E., Paulsen, I.T. & Salzberg, S.L. The value of complete microbial genome sequencing (you get what you pay for). *J. Bacteriol.* **184**, 6403–6405 (2002).
12. English, A.C. *et al.* Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology. *PLoS ONE* **7**, e47768 (2012).
13. Rasko, D.A. *et al.* Origins of the *E. coli* strain causing an outbreak of hemolytic-uremic syndrome in Germany. *N. Engl. J. Med.* **365**, 709–717 (2011).
14. Bashir, A. *et al.* A hybrid approach for the automated finishing of bacterial genomes. *Nat. Biotechnol.* **30**, 701–707 (2012).
15. Koren, S. *et al.* Hybrid error correction and *de novo* assembly of single-molecule sequencing reads. *Nat. Biotechnol.* **30**, 693–700 (2012).
16. Ribeiro, F.J. *et al.* Finished bacterial genomes from shotgun sequence data. *Genome Res.* **22**, 2270–2277 (2012).
17. Sommer, D.D., Delcher, A.L., Salzberg, S.L. & Pop, M. Minimus: a fast, lightweight genome assembler. *BMC Bioinformatics* **8**, 64 (2007).
18. Treangen, T.J., Sommer, D.D., Angly, F.E., Koren, S. & Pop, M. Next generation sequence assembly with AMOS. *Curr. Protoc. Bioinformatics* **33**, 11.8 (2011).
19. Blattner, F.R. *et al.* The complete genome sequence of *Escherichia coli* K-12. *Science* **277**, 1453–1462 (1997).
20. Kurtz, S. *et al.* Versatile and open software for comparing large genomes. *Genome Biol.* **5**, R12 (2004).
21. Tindall, B.J. *et al.* Complete genome sequence of *Meiothermus ruber* type strain (21^T). *Stand. Genomic Sci.* **3**, 26–36 (2010).
22. Han, C. *et al.* Complete genome sequence of *Pedobacter heparinus* type strain (HIM $762\text{-}3^T$). *Stand. Genomic Sci.* **1**, 54–62 (2009).
23. Alkan, C., Coe, B.P. & Eichler, E.E. Genome structural variation discovery and genotyping. *Nat. Rev. Genet.* **12**, 363–376 (2011).
24. Ariyadasa, R. & Stein, N. Advances in BAC-based physical mapping and map integration strategies in plants. *J. Biomed. Biotechnol.* **2012**, 184854 (2012).
25. Liu, G.E., Alkan, C., Jiang, L., Zhao, S. & Eichler, E.E. Comparative analysis of Alu repeats in primate genomes. *Genome Res.* **19**, 876–885 (2009).
26. Travers, K.J., Chin, C.S., Rank, D.R., Eid, J.S. & Turner, S.W. A flexible and efficient template format for circular consensus sequencing and SNP detection. *Nucleic Acids Res.* **38**, e159 (2010).
27. Ewing, B., Hillier, L., Wendl, M.C. & Green, P. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res.* **8**, 175–185 (1998).
28. Rieder, M.J., Taylor, S.L., Tobe, V.O. & Nickerson, D.A. Automating the identification of DNA variations using quality-based fluorescence re-sequencing: analysis of the human mitochondrial genome. *Nucleic Acids Res.* **26**, 967–973 (1998).
29. Loomis, E.W. *et al.* Sequencing the unsequenceable: expanded CGG-repeat alleles of the fragile X gene. *Genome Res.* **23**, 121–128 (2013).
30. Zhang, X. *et al.* Improving genome assemblies by sequencing PCR products with PacBio. *Biotechniques* **53**, 61–62 (2012).
31. Carneiro, M.O. *et al.* Pacific Biosciences sequencing technology for genotyping and variation discovery in human data. *BMC Genomics* **13**, 375 (2012).
32. Chain, P.S.G. *et al.* Genome project standards in a new era of sequencing. *Science* **326**, 236–237 (2009).
33. Murray, I.A. *et al.* The methylomes of six bacteria. *Nucleic Acids Res.* **40**, 11450–11462 (2012).

ONLINE METHODS

Materials. *E. coli* K-12 MG1655 (NCBI reference sequence NC_000913.2, genome size of 4,639,675 bp) was obtained from the American Type Culture Collection. DNA was prepared from this strain by Lofstrand Labs Limited. *M. ruber* DSM1279 (NC_013946.1) and *P. heparinus* DSM2366 (NC_013061.1) DNA samples were isolated from cultures originally obtained from DSMZ (German Collection of Microorganisms and Cell Cultures). Reference genomes had been generated at the Joint Genome Institute by Sanger sequencing (*P. heparinus*) and a combination of Sanger and 454 sequencing (*M. ruber*) as part of the GEBA (Genomic Encyclopedia of Bacteria and Archaea) project[21,22]. The BAC clone VMRC53-364D19 was selected from a BAC library constructed by C. Amemiya and corresponds to HapMap sample NA12878.

SMRTbell DNA template libraries of ~8- to 10-kb average insert size for the bacterial samples were prepared according to the manufacturer's specification, with G-tubes (Covaris) used for fragmentation. The BAC SMRTbell average template library was generated using a Covaris DNA miniTUBE (Blue for 220 series, Covaris), resulting in an average insert size of ~3 kb. SMRT sequencing was carried out on the PacBio RS according to standard protocols, with the XL binding kit used in conjunction with the C2 sequencing kit for *E. coli* (1 × 120-min acquisition mode), C2 chemistry for *M. ruber*, *P. heparinus* (1 × 90 min) and the BAC (2 × 45 min). All runs were carried out with diffusion-based loading and analyzed using the standard primary data analysis. Sanger validation sequencing was carried out by McLab.

Source codes for an implementation of HGAP and Quiver, data sets and additional documentation are available at http://www.pacbiodevnet.com/HGAP and http://www.pacbiodevnet.com/quiver. Visualizations were carried out using Tablet[34] and Gepard[35].

HGAP workflow details

Mapping. So that all continuous long reads (CLRs) from the sequencing data generated by the PacBio RS for HGAP are used, the longest reads are selected with a prespecified length cutoff to provide the seeds for constructing preassemblies. Typically, we target about 20× genome coverage of such seed sequences so that a sufficient amount of coverage of preassembled reads will be generated for the subsequent assembly. The preassembled reads are constructed by aligning all reads to each of the seed reads. Each read is mapped to multiple targeted seed reads using BLASR[36]. The number of sequence reads mapping to the seed reads is controlled by the "- bestn" parameter when BLASR is called for mapping. This number should be smaller than the total coverage of the seed sequences on the genome. If the "- bestn" number is too high, it is likely that reads from similar repeats will be mapped to each other, which could result in consensus errors. If the chosen "- bestn" number is too low, the quality of the preassembled read consensus may be decreased. The optimal choice might also depend on DNA fragment–library construction, which can affect the subread length distribution. Currently, we obtain good results by empirically choosing "- bestn" as 12 reads to map to the seed reads. Further study will allow a reasonable choice for optimized results.

Preassembly. A mini-assembly of reads that are mapped to seed reads is now performed. Because the seed sequences are long, such mini-assembly can be done simply by an alignment-and-consensus step to construct single 'preassembled reads' for each of the seed reads. In our current implementation for HGAP, we use PBDAG-Con for generating the consensus (https://github.com/PacificBiosciences/pbdagcon). PBDAG-Con uses an algorithm that is based on encoding multiple sequence alignments with a directed acyclic graph to find a best path for best consensus[37]. In general, an algorithm is needed that can generate good consensus by eliminating insertion and deletion errors in the raw sequences efficiently for this preassembly step[38,39]. There are multiple approaches that can achieve this goal. We find that using a graph to represent multiple sequence alignments (MSAs) is helpful to effectively remove random insertion and deletion errors for generating the consensus from the graph. For example, a long stretch of random insertions is typically an isolated path in the graph that can be eliminated easily when the optimized path is sought as the consensus. The principle of the preassembly step is illustrated in **Figure 1**. The full description of the algorithm used in PBDAG-Con is presented in **Supplementary Note 1**.

Certain sequencing artifacts are removed during this preassembly process: for example, spurious low-quality sequences and/or chimeric reads (**Supplementary Fig. 15**). Generally, this involves a few steps of trimming the alignment hits and removing the regions that do not have a sufficient number or quality of reads mapped to the seed reads (**Supplementary Fig. 15a**). Long chimeric sequences can easily confuse an assembly during the stage of constructing layouts, and they are typically removed as the first step within a genome assembler[40,41]. The preassembly step described here relies on consistent information contained in all reads. As these rare chimeric reads are generated randomly, the chimera junction in a read will be random, i.e., a chimeric seed read would have zero or low coverage through the chimeric junction when other reads are mapped to it, shown by example in **Supplementary Figure 15b**. Through this process, a preassembly that contains the whole chimeric sequence is avoided. Through the early detection and removal of such artifacts, i.e., during the preassembly step, the best preassembled reads without artifacts are sent to an assembler.

One can think of the preassembly process to combine the reads into preassembled reads as an assembly of the shorter reads, using the long seeds to provide the overlap and layout information. The preassembled read can be seen as the "contig" output of the traditional overlay-layout-consensus assembler. Therefore, existing modularized infrastructure for genome assemblies can be used for this preassembly step, for example, AMOS[18], to assemble the shorter reads to generate preassembled reads for subsequent full-genome assembly. The details on how to use AMOS for this preassembly step, and the comparison of the final assembly results, are provided in **Supplementary Note 1** and **Supplementary Table 1**, respectively.

Assembly. After the preassembly step, the resulting preassembled reads typically have read accuracies above 99% (a full evaluation of the preassembly accuracy is presented in the Results). Therefore, the preassembled reads can be easily fed into any assembler that can accommodate long-read inputs. Typically, assemblers that use the overlap-layout-consensus (OLC) strategy are better for assembling such long preassembled reads. In this study, we have used both the Celera Assembler[42] and MIRA (http://sourceforge.net/projects/mira-assembler/) for the

doi:10.1038/nmeth.2474

assembly process. For both assemblers, we found that the preassembled read accuracy was sufficiently high that we were able to use the parameters and configurations originally designed for Sanger sequencing reads. In this study, we did not generate quality values for preassembled reads. For assemblers that require quality values as input, we assigned a uniform phred score of QV 24 for all bases in the preassembled reads. We find that assigning such an *ad hoc* QV does not affect the assembly results, as the quality of preassembled reads is generally uniform because SMRT sequencing shows very little quality fluctuations through the reads[15]. In addition, the assembler is used only to generate the genome contiguity—the Quiver consensus algorithm (described in detail below) is used to determine the final genome sequence.

We have observed that sometimes an assembler will generate contigs that can be further connected. We have used minimus2 to connect them[17], although one should use caution and examine the result of this post-assembly step to ensure that contigs were combined properly and that no valid contigs were dismissed. In addition, we have noticed that sometimes contigs are supported by only a very small number of reads. A remapping process comparing the reads to these contigs can also be used to remove such artifacts from the assemblers used. In addition, the assembler will often generate overlapping ends on circular genomes that can be trimmed manually at the end of the assembly process. This was not done in this study so as to highlight the unaltered output of the assembler, and it explains the slightly larger numbers of assembled genome sizes relative to their references.

Final consensus. Upon obtaining the final result from the long-read assembler, we apply a new multiread consensus algorithm, called Quiver, to generate the best consensus sequence for the final genome sequence result. The Quiver algorithm is designed to take advantage of the full information from the raw pulse and base-call information that is generated during SMRT sequencing. During the signal processing, which converts the raw fluorescence pulses from a nucleotide incorporation event[43,44] into base calls, a hidden Markov model informs about the probabilities that these events corresponded to true incorporations or spurious base calls. The model is therefore specific to a particular SMRT sequencing chemistry and requires a training step. Quiver takes this QV-aware model of SMRT sequencing errors into account and uses a greedy algorithm to identify the maximum-likelihood consensus sequence corresponding to multiple reads.

The SMRT sequencing reads and the initial *de novo* assembly are the inputs to Quiver; the new consensus sequence and a table of corrected variants to the initial reference are its output. The algorithm first uses reference-based alignment to map the reads to their corresponding locations in an assembly. After reads are mapped to genomic regions, Quiver disregards the assembly sequence and the alignment. Thus, within a given genomic region, the consensus is computed anew from the reads alone, making it independent of fine-scale errors made in the draft assembly and free from local reference biases—as consensus bases have a tendency to agree with the reference owing to preferential mapping[31]. Quiver then uses a fast heuristic algorithm (partial order alignment[37]) to generate an initial, approximate consensus. All single-base substitution, insertion and deletion edits to the approximate consensus are tested, and those that improve the likelihood are applied, yielding an improved consensus. This procedure is repeated until the likelihood cannot be increased any further. Full details of this algorithm can be found in **Supplementary Note 1**.

34. Milne, I. *et al.* Tablet—next generation sequence assembly visualization. *Bioinformatics* **26**, 401–402 (2010).
35. Krumsiek, J., Arnold, R. & Rattei, T. Gepard: a rapid and sensitive tool for creating dotplots on genome scale. *Bioinformatics* **23**, 1026–1028 (2007).
36. Chaisson, M.J. & Tesler, G. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. *BMC Bioinformatics* **13**, 238 (2012).
37. Lee, C., Grasso, C. & Sharlow, M.F. Multiple sequence alignment using partial order graphs. *Bioinformatics* **18**, 452–464 (2002).
38. Edgar, R.C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).
39. Rausch, T. *et al.* A consistency-based consensus algorithm for *de novo* and reference-guided sequence assembly of short reads. *Bioinformatics* **25**, 1118–1124 (2009).
40. Huang, X. An improved sequence assembly program. *Genomics* **33**, 21–31 (1996).
41. Kelley, D.R., Schatz, M.C. & Salzberg, S.L. Quake: quality-aware detection and correction of sequencing errors. *Genome Biol.* **11**, R116 (2010).
42. Myers, E.W. *et al.* A whole-genome assembly of *Drosophila*. *Science* **287**, 2196–2204 (2000).
43. Eid, J. *et al.* Real-time DNA sequencing from single polymerase molecules. *Science* **323**, 133–138 (2009).
44. Korlach, J. *et al.* Real-time DNA sequencing from single polymerase molecules. *Methods Enzymol.* **472**, 431–455 (2010).