Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Microbiology

One chromosome, one contig: complete microbial genomes from long-read sequencing and assembly

Sergey Koren and Adam M Phillippy

Like a jigsaw puzzle with large pieces, a genome sequenced with long reads is easier to assemble. However, recent sequencing technologies have favored lowering per-base cost at the expense of read length. This has dramatically reduced sequencing cost, but resulted in fragmented assemblies, which negatively affect downstream analyses and hinder the creation of finished (gapless, high-quality) genomes. In contrast, emerging long-read sequencing technologies can now produce reads tens of kilobases in length, enabling the automated finishing of microbial genomes for under $1000. This promises to improve the quality of reference databases and facilitate new studies of chromosomal structure and variation. We present an overview of these new technologies and the methods used to assemble long reads into complete genomes.

Addresses
National Biodefense Analysis and Countermeasures Center,
110 Thomas Johnson Drive, Frederick, MD 21702, United States

Corresponding author: Phillippy, Adam M (phillippya@nbacc.net)

Current Opinion in Microbiology 2015, **23**:110–120

This review comes from a themed issue on **Genomics**

Edited by **Neil Hall** and **Jay CD Hinton**

For a complete overview see the Issue and the Editorial

Available online 1st December 2014

http://dx.doi.org/10.1016/j.mib.2014.11.014

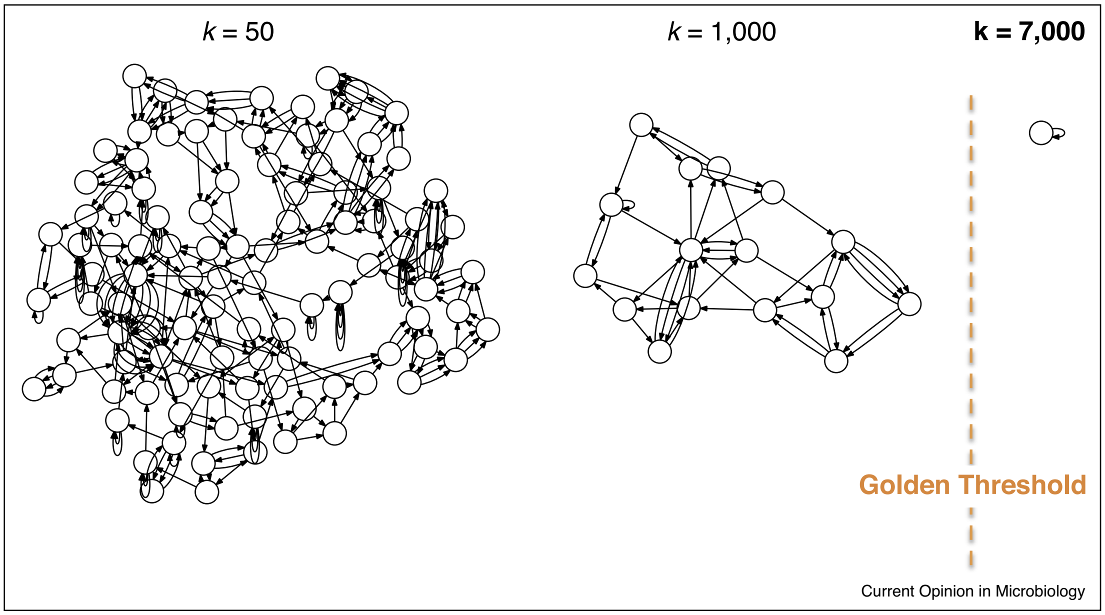
Introduction

The DNA sequencing revolution of the past decade has accelerated the study of microbial genomes, enabling projects on a scale previously unimaginable [1–16]. However, current large-scale studies have produced mostly short-read data (i.e. reads a few hundred bases in length) that are either mapped to a reference genome [4] or computationally assembled as a draft genome, forgoing the manual finishing process that was previously standard [17]. Between 2007 and 2011, the percentage of genomes being finished to remove all gaps and ensure a per-base accuracy of >99.99% was below 35% [18••,19]. This has lowered the average quality of genomes in the public databases, limiting the types of analyses that can be performed. For example, studies of structural variation are severely limited by the use of short reads, which cannot resolve large-scale structural mutations mediated by repetitive transposable elements. While short reads are sufficient for many other analyses, such as strain typing, outbreak tracing, and pan-genome surveys, the accuracy of these studies is improved by the inclusion of finished genomes. Unfinished genomes can lead to mapping artifacts, missed gene calls, and inaccurate repeat construction. However, due to the historically higher cost of finished genomes, recent studies have almost exclusively produced unfinished genomes — relying on just a few finished references for annotation and mapping. Long-read sequencing, with read lengths on the order of ten thousand bases, promises to change this cost versus quality trade-off by enabling the complete assembly and economical finishing of microbial genomes.

Genome assembly reconstructs a genome from many shorter sequencing reads [20–22]. The primary challenge to all assembly algorithms is repeats, which can be resolved using either long reads or DNA inserts (e.g. fosmids) [23]. However, while sequencing throughput has dramatically increased in past years, read length has remained relatively limited, leaving most genomes fractured into hundreds of pieces. Assemblers typically represent these uncertain reconstructions as graphs of contigs (Figure 1a). To help resolve repeats and improve the reconstruction, it is possible to sequence paired reads. *Paired ends* refer to the sequenced ends of larger, size-selected inserts, which are separated by a known distance within some error bound. This information serves as a secondary constraint on assembly that can resolve repeats and improve continuity. However, due to uncertainty in the size selection process and the unknown sequence between the ends, this technique cannot resolve all repeat types, leaving some regions of the genome uncharacterized [24,25]. In contrast, merely increasing read length can significantly simplify the assembly problem [26] and resolve more repeats (Figure 1b and c).

Two types of repeats complicate assembly, global and local. An example of a global repeat is the prokaryotic rDNA operon, which is typically less than 7 kilobase pairs (7 kbp) in length [27]. This type of repeat consists of a long sequence duplicated throughout the genome, which tangles the assembly graph and creates ambiguous adjacency relationships between contigs. In contrast, a local repeat typically comprises a simple sequence unit, sometimes only a few base pairs in length, which is repeated in tandem many times. A common example in prokaryotic genomes are variable number tandem repeats (VNTRs)

Figure 1



Read length simplifies the assembly problem. Simplified assembly graphs for the *Escherichia coli* K-12 genome are shown for varying read lengths *k* [26]. Contigs are nodes and unresolved adjacencies are represented as links. **(a)** With a short read length, many regions of the genome remain unresolved. **(b)** Increasing the length to 1 kbp, similar to that produced by Sanger sequencing [84], resolves a large fraction of the genome, but some complexity remains due to large global repeats. **(c)** Once the length is above the 'golden threshold,' which exceeds the most common repeat length in prokaryotic genomes, all ambiguity is removed from the graph (this figure was derived from the same data used in Figure 1 of [18••]).

[28]. Local repeats introduce local complexity, often exhibited as short cycles in the assembly graph. These repeats, though often shorter than global repeats, have traditionally been difficult to resolve, because they appear as arrays of multiple elements. If the array is longer than the read length, it is not always possible to determine the correct copy number. Typically, paired-ends cannot resolve this type of repeat because the element size is smaller than the uncertainty of the distance estimate, making confident reconstruction impossible. Thus, fully resolving a prokaryotic genome requires both long-range linking to resolve global repeats and base-pair resolution to resolve local repeats. Long reads provide both of these characteristics.

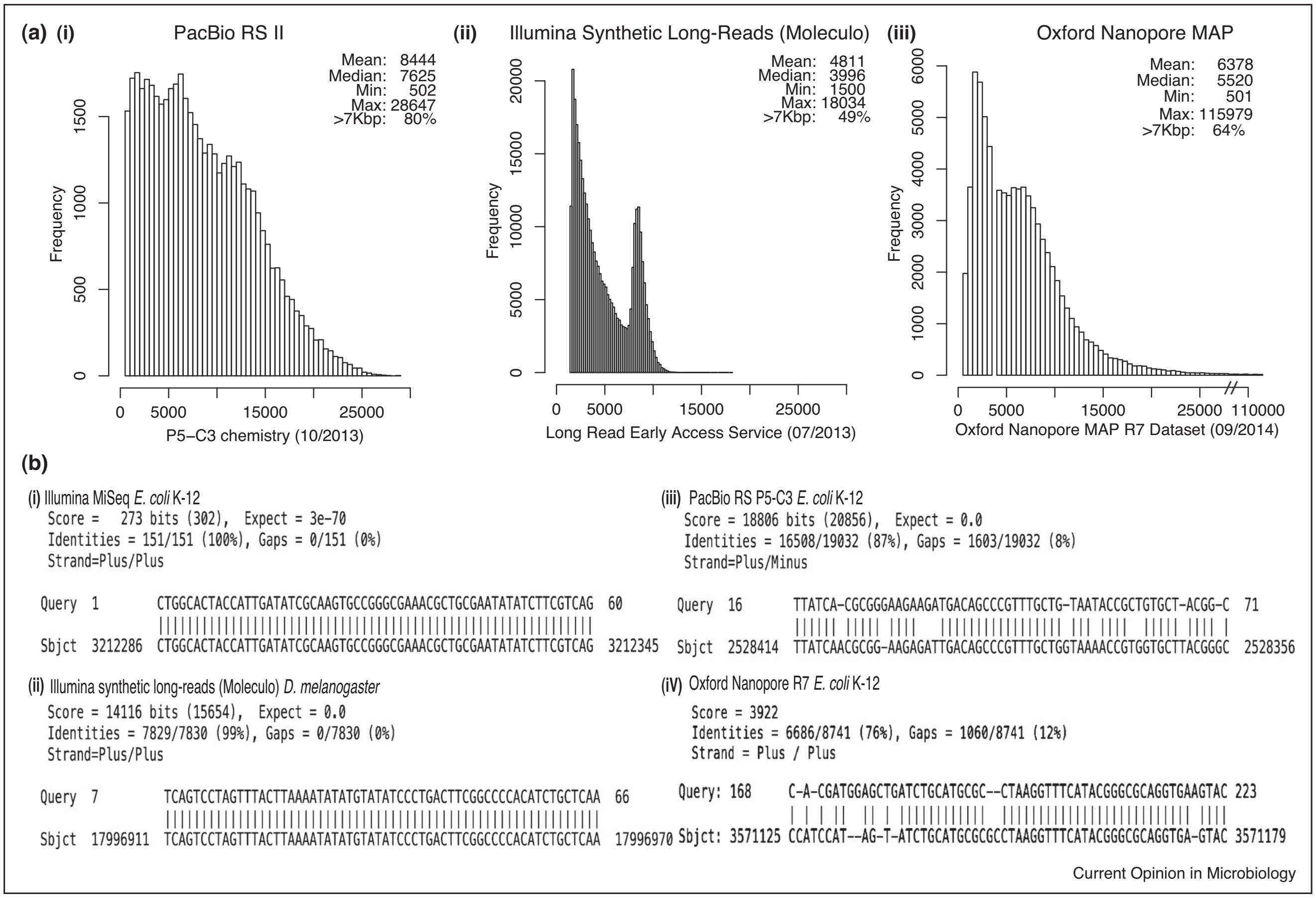
On the basis of an analysis of repeat size and count of all sequenced repeats, we previously classified microbial genomes into three complexity classes [18••]. The rDNA operon (between 5 and 7 kbp in length) was found to be the most prevalent large repeat and was used as the basis for measuring complexity. The first two classes (I and II), where the largest repeat is the rDNA operon, comprise approximately 77% of currently finished genomes [18••]. The remaining genomes, with at least one repeat >7 kbp, were defined as class III. Thus, a sequencing technology that can generate sequences over the 7 kbp 'golden threshold' (Figure 1c) would be expected to resolve almost 80% of known microbial genomes. Until recently, 7 kbp was more than ten times longer than typical sequencing lengths. However, new sequencing technologies have recently exceeded this threshold and, for the first time, enabled the complete assembly of most microbial genomes without gaps [18••,29••,30••].

There are currently three technologies capable of long-read sequencing (defined here as producing reads greater than 7 kbp). Below, we outline these three technologies, approaches to utilize them, and remaining challenges. As of writing, the PacBio RS instrument has been the most widely used of the three, and the only to demonstrate assembly of complete genomes, so the later sections will focus on this instrument. We conclude with current recommendations for sequencing and assembling microbial genomes.

PacBio RS

The PacBio RS was the first commercially available long-read sequencer, capable of producing kilobase-sized reads with an average accuracy of ~82% upon release in 2011 [31,32•]. The technology is based on single molecule, real-time (SMRT) sequencing, which uses zero-mode waveguides to observe the base incorporations of an anchored polymerase. The instrument preferentially outputs a large number of relatively short sequences

Figure 2



(i) Illumina MiSeq *E. coli* K-12

```
  Score =   273 bits (302),  Expect = 3e-70
  Identities = 151/151 (100%), Gaps = 0/151 (0%)
  Strand=Plus/Plus

Query  1        CTGGCACTACCATTGATATCGCAAGTGCCGGGCGAAACGCTGCGAATATATCTTCGTCAG  60
                ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  3212286  CTGGCACTACCATTGATATCGCAAGTGCCGGGCGAAACGCTGCGAATATATCTTCGTCAG  3212345
```

(ii) Illumina synthetic long-reads (Moleculo) *D. melanogaster*

```
  Score = 14116 bits (15654),  Expect = 0.0
  Identities = 7829/7830 (99%), Gaps = 0/7830 (0%)
  Strand=Plus/Plus

Query  7         TCAGTCCTAGTTTACTTAAAATATATGTATATCCCTGACTTCGGCCCCACATCTGCTCAA  66
                 ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  17996911  TCAGTCCTAGTTTACTTAAAATATATGTATATCCCTGACTTCGGCCCCACATCTGCTCAA  17996970
```

(iii) PacBio RS P5-C3 *E. coli* K-12

```
  Score = 18806 bits (20856),  Expect = 0.0
  Identities = 16508/19032 (87%), Gaps = 1603/19032 (8%)
  Strand=Plus/Minus

Query  16       TTATCA-CGCGGGAAGAAGATGACAGCCCGTTTGCTG-TAATACCGCTGTGCT-ACGG-C  71
                |||||| ||||| ||||   ||||||||||||||||| ||| ||||  ||||| |||| |
Sbjct  2528414  TTATCAACGCGG-AAGAGATTGACAGCCCGTTTGCTGGTAAAACCGTGGTGCTTACGGGC  2528356
```

(iV) Oxford Nanopore R7 *E. coli* K-12

```
  Score = 3922
  Identities = 6686/8741 (76%), Gaps = 1060/8741 (12%)
  Strand = Plus / Plus

Query: 168     C-A-CGATGGAGCTGATCTGCATGCGC--CTAAGGTTTCATACGGGCGCAGGTGAAGTAC 223
               | | | ||  || | ||||||||||||  |||||||||||||||||||||||||| ||||
Sbjct: 3571125 CCATCCAT--AG-T-ATCTGCATGCGCGCCTAAGGTTTCATACGGGCGCAGGTGA-GTAC 3571179
```

Length distributions and example alignments for current long-read sequencers. **(a)** The lengths produced from current iterations of long-read sequencers. In all cases only sequences >500 bp are included. (i) The PacBio RSII read distribution is based on a single SMRTcell of *E. coli* K-12 filtered subreads using the P5C3 chemistry. (ii) The Moleculo length distribution is based on a publically released dataset from *D. melanogaster* [85]. (iii) The Oxford Nanopore MinION Access Program provides customers with early-access instruments. The length distribution is given for the most recent public data at the time of writing [45]. Because this is publically sourced data it is uncontrolled for library preparation methods, but initial reports indicate that MinION read lengths mirror the lengths of the molecules in the sample [86]. **(b)** Example alignments of three long-read technologies as well as a short-read technology. (i) Illumina MiSeq alignment. (ii) Moleculo read alignment with high identity. (iii) PacBio RSII P5C3 chemistry alignment. (iv) Oxford Nanopore R7 2D sequence alignment [43]. Note, this is a '2D' read for which both strands of a double-stranded molecule were read, which represents the upper limit of current MinION accuracy. Web BLAST [87] was run with default options for 'Somewhat similar sequences' in all cases except for Oxford, which was aligned with last [88]: lastal -r 1 -q 1 -a 1 -b 1 -Q 0 -j 3 -f 1 ecoli. The output was converted to BLAST format using maf-convert.py and a randomly selected sequence aligned over at least 90% of its length was selected for display.

and a diminishing number of long sequences, producing a log-normal length distribution (Figure 2ai) [33]. Due to the initially high error rate and relatively short average read length, the first SMRT reads could only be used in combination with other complementary, high-accuracy technologies [30••,34••,35••].

However, rapid advances in chemistry and library preparation have boosted both the median and maximum sequence lengths [18••,33] to 10 kbp and 50 kbp respectively in the latest runs [36••,37••]. The average per-read accuracy has also increased to ~87% [18••]. This has enabled non-hybrid assembly of SMRT reads [18••,29••]. In addition to resolving global repeats, SMRT sequencing exhibits relatively little sequencing bias [34••,38], allowing the resolution of low complexity sequence (e.g. high/low %GC) and correction of per-read errors using statistical methods to achieve finished-grade consensus sequence (>99.99% accurate). These characteristics have led to the first automated assemblies of finished bacterial genomes [18••,29••].

Illumina synthetic long reads, Moleculo

Illumina synthetic long-read sequencing, previously known as Moleculo [39,40], relies on an advanced library preparation technique to pool barcoded subsets of the genome, allowing the construction of synthetic long reads [41]. The resulting synthetic reads are extremely high quality (>99% base accuracy) [40] but currently limited to approximately 18 kbp in length [41]. Since the construction of the synthetic sequences relies on local assembly, it is also limited by local repeat structure that cannot be resolved by the underlying pooled data. This produces a bi-modal read distribution, with a peak of long sequences 8–10 kbp and a mass of shorter sequences that are broken by repeat structure or library preparation (Figure 2aii). In addition, the Illumina sequencing process has known biases that can result in coverage gaps, especially in extreme %GC regions [38]. In a direct comparison, PacBio better resolved transposable elements and produced contigs over 400 fold longer than Moleculo for the assembly of *Drosophila melanogaster* [36••]. Thus, Moleculo may see limited use for *de novo* assembly, and is more likely to be useful for haplotype phasing and metagenomic applications.

Oxford Nanopore MinION

The most recent, and potentially most disruptive long-read technology to emerge is based on nanopore sequencing. Rather than relying on optics, like the PacBio and Illumina systems, the MinION is a thumb-drive sized device that measures deviations in electrical current as a single DNA strand is passed through a protein nanopore [42]. The size, robustness, and affordability of the MinION make it entirely unique, and more akin to a mobile sensor than a traditional sequencer. At the time of writing, the MinION is only available via invitation and little data has been publically released, but early results show the instrument can produce sequences well above the length threshold required for microbial finishing (Figure 2aiii) [43–45]. However, the initially reported accuracy (∼70% for the R7 chemistry and ∼80% for R7 2D sequences [45]) is significantly lower than alternatives from PacBio and Illumina. In addition, these accuracy estimates are based only on regions of the reads that were successfully aligned, suggesting the average identity across the entire read length is even less. This presents a problem for assembly, and it is unclear if current *de novo* methods can handle the increased error. Thus, initial attempts to assemble MinION data are likely to mirror the hybrid approaches developed for early SMRT sequencing, and non-hybrid assembly of nanopore data will require technological advancements and/or improved algorithms for base calling and alignment. It is likely that all tools previously developed for SMRT sequencing will require modification to handle the specific characteristics of nanopore data.

Algorithms

Being the first technology to market, several algorithmic approaches have been developed to handle PacBio SMRT sequencing data. Though developed for SMRT reads, these same strategies are likely to apply to future long-read technologies. Each of the approaches has strengths and weaknesses and no single one is best for all applications. Table 1 gives a summary of available methods, a description, and the minimum long-read coverage required to operate (ranging from 50× to <10×). However, as all long-read technologies produce a distribution of sizes (Figure 2), the ability to fully assemble a genome depends entirely on collecting sufficient long-read coverage to resolve repeats. Redundant coverage is required to increase the likelihood that at least one long read spans, from end to end, each repetitive element in the genome. Thus, even algorithms capable of working with low-coverage data benefit from higher coverage [46•]. Generating sufficient data over the golden threshold previously required as much as 200× coverage of SMRT sequencing [18••]. But, as a larger fraction of long reads are generated, this coverage requirement decreases rapidly [18••], with current SMRT chemistries requiring only 50× for finished assemblies (approximately a single PacBio SMRTcell for an average bacterium). Below we outline the current strategies for assembling long reads for various coverage and data scenarios.

Traditional OLC (Allora [47•], Celera Assembler [48])

Analogous to the method first used for whole-genome sequencing [48], an assembly is constructed directly from long reads using an Overlap-Layout-Consensus (OLC) approach. This strategy can be readily applied to accurate long reads, like Moleculo. However, detecting sequence overlaps at the error rates produced by single-molecule sequencing instruments is computationally expensive and introduces false-positive and false-negative overlaps. These erroneous overlaps complicate the assembly graph and can lead to fragmented or mis-assembled genomes. This has, to date, limited the use of the traditional approach to high-accuracy reads. Although, OLC assembly is commonly used to assemble corrected reads using the hierarchical methods described below.

Hierarchical (Figure 3a)

Hierarchical hybrid (PBcR [34••], LSC [49•], ECTools [37••], LoRDEC [50•], proovread [51•], DBG2OLC [52•])

To address the difficulty of building an overlap graph directly from noisy reads, the hierarchical approach first improves the quality of the long reads in a process called correction, scrubbing, or preassembly. As introduced by Koren *et al.* [34••], long-read correction involves mapping multiple reads to a single long read to identify and correct errors using a consensus alignment. In hybrid mode, this approach uses a complementary technology (such as Illumina short reads) to correct the long, noisy sequences. The corrected sequences are highly accurate and can be assembled using a traditional OLC approach [48,53].

More recent approaches have focused on performance improvements from preassembling the secondary technology prior to correction [37••], alignment to a *de Bruijn* graph [50•], or other applications of long reads such as

Table 1

Summary of software tools for long-read assembly

Name	Description	Website	Minimum coverage[a]
Celera Assembler [48]	An archetypical OLC assembler originally designed for Sanger data and used to assemble the first human genome. Support has been subsequently added for 454 [78] and long reads [34••]. The majority of hierarchical methods rely on Celera Assembler for assembling corrected reads. The most recent version adds unpublished support for assembly of lower-coverage raw PacBio data	http://wgs-assembler.sourceforge.net/	PacBio 20×
Allora [47•]	An OLC assembler built on the AMOS [79,80] infrastructure. Can work on raw PacBio data. It is no longer maintained and was only recommended for small genomes	N/A	Unknown
Falcon	An experimental OLC assembler designed to preserve ambiguity in the assembly graph. While most assemblers will break an assembly at a haplotype boundary, Falcon will output the longest path through the graph along with alternate paths. This is similar to approaches previously applied to metagenomes [81,82]. Currently, it can be used only with high-accuracy corrected sequences	https://github.com/PacificBiosciences/FALCON	PacBio 10×
DBG2OLC [52•]	A hybrid method with a focus on computational efficiency. The algorithm identifies PacBio sequences that could be used to improve Illumina contigs. The Illumina contigs are used to identify PacBio overlaps which are then used to generate an OLC assembly	https://sites.google.com/site/dbg2olc/	PacBio 10× Illumina 50×
ECTools [37••]	A hierarchical hybrid method that uses pre-assembled data rather than raw sequences for correction. The pre-assembled data improves both efficiency and quality of long read correction. After correction, sequences are assembled using Celera Assembler	https://github.com/jgurtowski/ectools	PacBio 20× Illumina 50×
LoRDEC [50•]	An approach combining algorithms from hierarchical hybrid and read threading. This method aligns long reads to a *de Bruijn* graph and generates corrected sequence for long reads by traversing paths in the graph	http://www.atgc-montpellier.fr/lordec/	Unknown
LSC [49•]	A hierarchical hybrid method targeted to RNA and cDNA sequences. It first compresses the short and long sequences to remove homopolymers and then aligns and corrects the long read sequences	http://www.healthcare.uiowa.edu/labs/au/LSC/	Unknown
proovread [51•]	A hierarchical hybrid method that utilizes an iterative strategy to accelerate correction. Sequences are first mapped at moderate sensitivity and successfully corrected regions masked during subsequent iterations	http://proovread.bioapps.biozentrum.uni-wuerzburg.de/	Unknown
PBcR [18••,34••]	The original hierarchical method. First designed for hybrid assembly, recent versions support both hybrid and non-hybrid approaches. After correction, sequences are assembled using Celera Assembler. It has demonstrated finished bacterial genomes from both hybrid and non-hybrid data as well as high-quality eukaryotic genomes	http://wgs-assembler.sourceforge.net/wiki/index.php?title=PBcR	PacBio 50× or PacBio 20× Illumina 50×
HGAP [29••]	A hierarchical non-hybrid method similar to PBcR and the first to demonstrate non-hybrid assembly. After correction, sequences are assembled using Celera Assembler. It has demonstrated finished bacterial genomes using high-coverage long reads as well as high-quality eukaryotic genomes	https://github.com/PacificBiosciences/Bioinformatics-Training/wiki/HGAP-in-SMRT-Analysis	PacBio 50×
Dazzler [55••]	A non-hybrid hierarchical assembler focusing on efficiency both in speed and disk usage. As of writing, the full assembler has not yet been publically released	http://dazzlerblog.wordpress.com/	Unknown

Table 1 (*Continued*)

Name	Description	Website	Minimum coverage[a]
Sprai [56•]	A non-hybrid method similar to HGAP and PBcR. Sequences are first corrected and then assembled using Celera Assembler. It has been used to finish at least one genome	http://zombie.cb.k.u-tokyo.ac.jp/sprai/index.html	PacBio 50×
Allpaths-LG [30••]	The original long read threading method. A *de Bruijn* assembler originally designed for large genomes, it requires both short overlapping Illumina sequences as well as long-range pairs. Support was added for long reads, which are used to resolve local repeat structure in the assembly graph. It has demonstrated high-quality finished bacterial genomes given a mix of short inserts, long inserts, and long read data	http://www.broadinstitute.org/software/allpaths-lg/blog/	PacBio 10× Illumina 50×
SPAdes [62]	A *de Bruijn* assembler originally designed for single-cell sequencing data, it has since been shown to work well on microbial assembly [25]. The latest release has added support for long reads and the documentation indicates it can produce finished bacterial genomes given sufficient long-read coverage	http://bioinf.spbau.ru/spades	PacBio 10× Illumina 50×
Cerulean [63•]	An assembly boosting method that utilizes an assembly graph generated by ABySS [83]. Long reads are aligned to contigs to generate linking information. The original assembly graph and long-read linking information are analyzed to assemble scaffolds	https://sourceforge.net/projects/ceruleanassembler/	PacBio 10× Illumina 50×
PBJelly [35••]	A gap-filling approach that takes scaffolds generated by any assembler and fills scaffold gaps using long reads. Originally, only internal scaffold gaps could be resolved, but a recent version has added support for merging multiple scaffolds	https://sourceforge.net/projects/pb-jelly/	PacBio 5× Illumina 50×
AHA [59••]	The original long read scaffolding method. AHA uses contigs from any assembler and converts long read alignment data into assembly constrains. It relies on the Bambus [60] scaffolder to produce scaffolds	https://github.com/PacificBiosciences/Bioinformatics-Training/wiki/AHA	PacBio 5× Illumina 50×
SSPACE-LongRead [46•]	A recently published scaffolding approach that can scaffold contigs from any assembler. Single-chromosome assembly was demonstrated on at least one bacterial genome and presented results were competitive with AHA	http://www.baseclear.com/lab-products/bioinformatics-tools/sspace-longread/	PacBio 5× Illumina 50×

[a] Minimums reflect sufficient coverage for the tool to operate as intended. Minimum coverage also depends on read length and quality. Assembly results will not be equal between all tools at these values, and the performance of most tools will improve with additional coverage. Regardless of the tool used, microbial genome finishing with PacBio P5C3 currently requires a minimum of ~50× coverage.

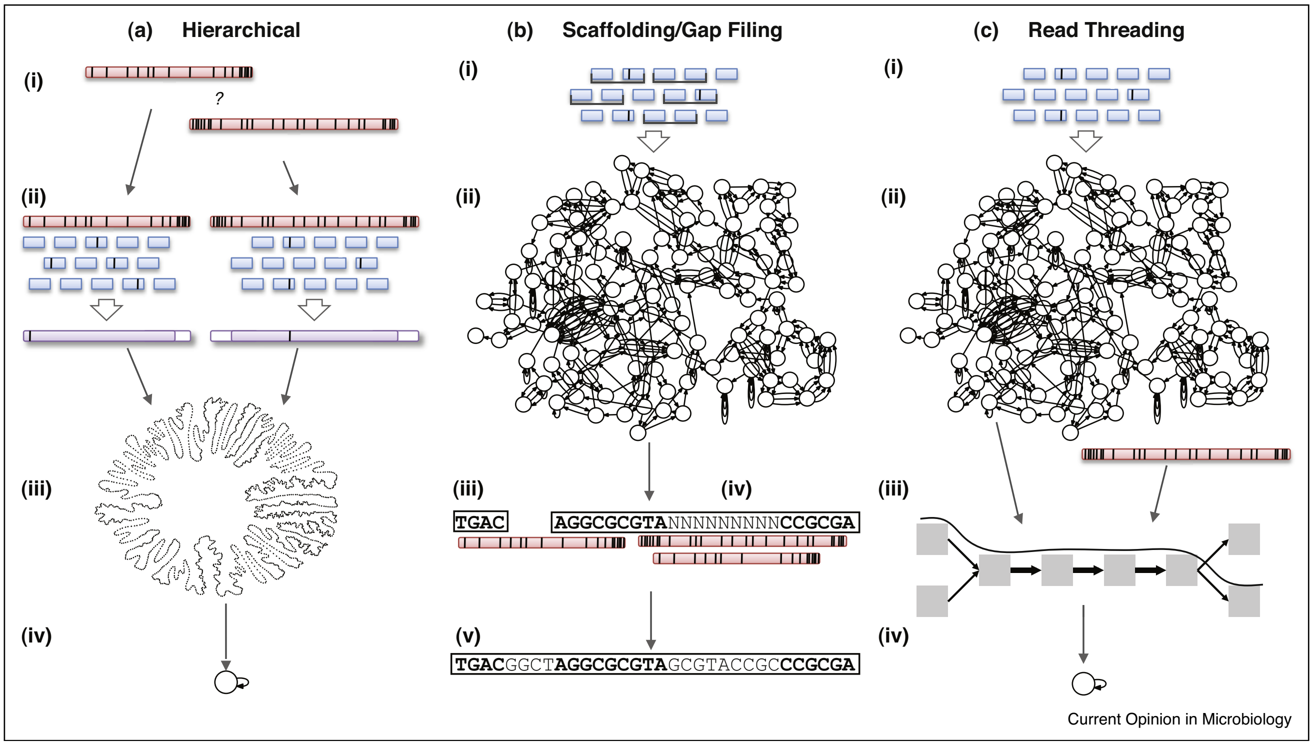
RNA-Seq [49•]. These algorithms are more efficient and can work well with lower coverage (20–50×), but due to systematic biases in amplification-based sequencing and mapping error, non-hybrid methods begin to outperform hybrid methods for higher long-read coverage [18••,37••].

Hierarchical non-hybrid (PBcR [18••], HGAP [29••], Dazzler [54,55••], Sprai [56•])

As with hierarchical hybrid approaches, these algorithms first correct the long reads. However, rather than align a secondary technology, the long reads are aligned only against each other. The most obvious overlaps are identified first, usually consisting of one read entirely contained within the length of another. This effectively clusters shorter reads with the long reads that contain them. The longest sequences are then corrected using a consensus of the data and assembled with an OLC method. As SMRT sequencing improved to longer and more accurate sequences, this became a highly effective strategy for assembly [18••,57]. The resulting contigs can be further polished using a quality-aware statistical method [29••] to obtain extremely accurate results [18••].

Hierarchical non-hybrid algorithms require relatively high levels of coverage (>50×) and have required as much as 24 h to assemble a bacterial genome. However, they have been proven to generate highly accurate, finished genomes. Recent algorithm development has focused on improving the efficiency of this process [36••,54,58] — lowering runtime for bacterial assembly to under an hour on a single server [36••].

Figure 3



Overview of long read assembly methods. **(a)** The hierarchical method uses a two-step process. Due to the difficulty of detecting noisy overlaps (i), other sequences (ii) are used to correct the noisy data. The sequences used for correction can be high-identity reads, assemblies of other reads, or even other long reads. Once the noisy reads are corrected, an assembly graph (iii) is constructed from the now high-accuracy data and the graph is simplified to a single contig (iv). **(b)** The scaffolding and gap patching methods start with complementary high-identity sequence data, which is first assembled. Once assembled, long reads are aligned and used to connect contigs (iii) or fill in missing sequence (iv). The filled and scaffolded contigs are output (v). **(c)** Read threading relies on resolving a short-read assembly graph using long reads. It differs from the scaffolding approach because it operates on the assembly graph rather than the contigs. First, an assembly graph is constructed from the complementary data (ii). The long reads are then layered onto the graph (iii) and used to resolve repeat-induced graph structures. The graph structure may then be resolved (iv).

Assembly boosting (Figure 3b,c)

Gap-filling (PBJelly [35••])

Gap-filling works with an existing assembly to fill in missing sequence [35••]. This approach has the advantage of not removing or breaking existing contigs, making it easy to transfer features to the new assembly. This makes gap-filling well suited for assembly improvement, where there exists a trusted assembly with large scaffolds. The long read data is layered on top of the existing assembly to close scaffold gaps. In addition, gap-filling can be performed with low coverage data (<10×), making it an economical option. However, closure of all gaps is rarely achieved with this approach.

Scaffolding (AHA [59••], SSPACE-LongRead [46•])

Scaffolding also works on an existing assembly, but focuses on joining, ordering, and orienting contigs connected by long reads [59••]. The approach originally relied on pre-existing scaffolding tools [60], but newer tools have been developed that take better advantage of the long reads [46•]. As with gap-filling, these approaches work with low coverage data (<10×), but performance increases with higher coverage. Filling can be used to close gaps remaining after scaffolding. However, scaffolding can suffer from overly aggressive joining [61], especially when presented with mis-assembled contigs, so it is important to start with a highly accurate assembly when scaffolding with long reads.

Read threading (Allpaths-LG [30••], SPAdes [62], Cerulean [63•])

Read threading is differentiated from the other two boosting approaches because it operates directly on the assembly graph. This can result in more accurate and complete assemblies. The read threading problem, or Eulerian Superpath Problem as described by Pevzner *et al.* [64], is to find a path through the *de Bruijn* graph that is consistent with the sequencing reads. In the general case, the read threading problem is NP-hard and cannot be optimally solved [65], but heuristics can be effective in practice. This allows for the resolution of repeats shorter than the read length (see Miller *et al.* [20]

for a review). The first threading method developed for long reads used them to resolve local repeats, while long-range pairs were used to resolve global repeats. This approach demonstrated near-finished genomes of extremely high quality [30••]. More recent approaches have relied on long reads to resolve repeats of all types [62,63•]. As with the other assembly boosting techniques, read threading has no minimum coverage requirement (<10×).

Discussion

PacBio SMRT sequencing has, for the first time, enabled automated and complete assembly of microbial genomes. This has been demonstrated by multiple methods on multiple genomes [18••,29••,30••], and independent studies have confirmed that the resulting assemblies are of finished quality (no gaps and >99.99% accurate) [66–69]. Assembly comparisons have demonstrated that the hierarchical non-hybrid approach outperforms others when sufficient coverage is available [18••,37••,61,70]. Increasing read lengths will increase the fraction of genomes that can be fully resolved using this approach. Emerging long-read technologies, such as nanopore sequencing, are expected to follow a similar development path, and may represent an additional option for genome finishing in the future.

Due to the uneven length distribution produced by SMRT sequencing, obtaining sufficient data to resolve repeats within a microbial genome typically requires high coverage. Therefore, 100× SMRT sequencing is currently the most reliable method for generating finished microbial genomes. In addition, careful DNA extraction and library preparation are crucial to isolate and sequence the longest molecules possible. However, size selection can inadvertently exclude short plasmids, so a secondary, short-fragment library may be required. At current prices, 100× coverage for an average-sized bacterial genome costs less than $1000 using a PacBio RSII with a 20 kbp library preparation [18••]. However, this cost fluctuates based on sequencing yield and genome complexity, with large class III genomes being the most costly. Independent sequencing providers are available for smaller laboratories that cannot justify the high instrument cost and prefer to pay per run. Also, for those without sufficient computing resources, a bacterial genome can now be assembled on a commodity cloud system, such as Amazon Web Services, for under $5 [71].

The preferred assembly method depends on the data collected. With at least 50× of SMRT data, hierarchical non-hybrid assembly (Section 'Hierarchical non-hybrid (PBcR [18••], HGAP [29••], Dazzler [54,55••], Sprai [56•])') is recommended. This method requires only a single sequencing library, typically exceeds the requirement for 'finished quality' (>99.99% accurate), and has been shown to outperform hybrid approaches with similar coverage [18••,37••]. If additional validation is desired, a SMRT assembly can be polished with short reads using a tool such as Pilon [72] and/or structurally validated with optical mapping [73,74]. However, these validation methods add cost and can be forgone in most cases. Although hybrid approaches can operate with less long-read data, this increases the likelihood of residual gaps. Thus, these approaches are typically reserved for the improvement of existing short-read assemblies via scaffolding (Section 'Scaffolding (AHA [59••], SSPACE-LongRead [46•])') or read threading (Section 'Read threading (Allpaths-LG [30••], SPAdes [62], Cerulean [63•])'). Such hybrid methods may also be preferable for future nanopore data, if throughput and accuracy remain limited. Finally, in the case of class III genomes, additional long-range linking information (e.g. an optical map) may be required to resolve the longest repeats and close all gaps.

With the recent availability of long reads, new challenges have also become apparent. One challenge is diploid (or other non-haploid) genomes and the full reconstruction of divergent alleles. While eukaryotic genomes can also be assembled using long reads, it is not currently possible to span large centromeres and fully resolve all chromosomes. Other techniques, such as chromatin interaction mapping with Hi-C [75–77], may help bridge this gap between long read assembly and large, complex genomes. A second, but related, challenge is characterizing and algorithmically handling the variation present in microbial populations. This applies to both metagenomic samples and clonal microbial cultures. Long reads reveal structural context and linked variants that go undetected by short reads, raising the possibility of separating individual genomes from a metagenome with long read assembly. Characterizing (and representing) the structural and allelic variation of microbial populations is an open problem that can now be addressed through the use of long-read sequencing. In addition, the decreased cost of finishing is expected improve the quality of reference databases and accelerate the study microbial chromosome structure, repetitive elements, and population variation.

Competing interests

None declared.

Acknowledgements

The authors would like to thank Joshua Quick, Aaron Quinlan, Nicholas Loman, Konrad Paszkiewicz, Audrey Farbos, Karen Moore, and Paul O'Neill for posting *E. coli* K-12 data and the first full sequencing runs from the Oxford Nanopore MAP program. The contributions of SK and AMP were funded under Agreement No. HSHQDC-07-C-00020 awarded by the Department of Homeland Security Science and Technology Directorate (DHS/S&T) 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 U.S. 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.

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Croucher NJ, Harris SR, Fraser C, Quail MA, Burton J, van der Linden M, McGee L, von Gottberg A, Song JH, Ko KS *et al.*: **Rapid pneumococcal evolution in response to clinical interventions**. *Science* 2011, **331**:430-434.

2. Snitkin ES, Zelazny AM, Thomas PJ, Stock F, Group NCSP, Henderson DK, Palmore TN, Segre JA: **Tracking a hospital outbreak of carbapenem-resistant *Klebsiella pneumoniae* with whole-genome sequencing**. *Sci Transl Med* 2012, **4**:148ra116.

3. Bryant JM, Grogono DM, Greaves D, Foweraker J, Roddick I, Inns T, Reacher M, Haworth CS, Curran MD, Harris SR *et al.*: **Whole-genome sequencing to identify transmission of *Mycobacterium abscessus* between patients with cystic fibrosis: a retrospective cohort study**. *Lancet* 2013, **381**: 1551-1560.

4. Comas I, Coscolla M, Luo T, Borrell S, Holt KE, Kato-Maeda M, Parkhill J, Malla B, Berg S, Thwaites G *et al.*: **Out-of-Africa migration and Neolithic coexpansion of *Mycobacterium tuberculosis* with modern humans**. *Nat Genet* 2013, **45**: 1176-1182.

5. Croucher NJ, Finkelstein JA, Pelton SI, Mitchell PK, Lee GM, Parkhill J, Bentley SD, Hanage WP, Lipsitch M: **Population genomics of post-vaccine changes in pneumococcal epidemiology**. *Nat Genet* 2013, **45**:656-663.

6. Eyre DW, Cule ML, Wilson DJ, Griffiths D, Vaughan A, O'Connor L, Ip CL, Golubchik T, Batty EM, Finney JM *et al.*: **Diverse sources of *C. difficile* infection identified on whole-genome sequencing**. *N Engl J Med* 2013, **369**:1195-1205.

7. Holden MT, Hsu LY, Kurt K, Weinert LA, Mather AE, Harris SR, Strommenger B, Layer F, Witte W, de Lencastre H *et al.*: **A genomic portrait of the emergence, evolution, and global spread of a methicillin-resistant *Staphylococcus aureus* pandemic**. *Genome Res* 2013, **23**:653-664.

8. Mather AE, Reid SW, Maskell DJ, Parkhill J, Fookes MC, Harris SR, Brown DJ, Coia JE, Mulvey MR, Gilmour MW *et al.*: **Distinguishable epidemics of multidrug-resistant Salmonella Typhimurium DT104 in different hosts**. *Science* 2013, **341**: 1514-1517.

9. Miotto O, Almagro-Garcia J, Manske M, Macinnis B, Campino S, Rockett KA, Amaratunga C, Lim P, Suon S, Sreng S *et al.*: **Multiple populations of artemisinin-resistant *Plasmodium falciparum* in Cambodia**. *Nat Genet* 2013, **45**:648-655.

10. Sheppard SK, Didelot X, Meric G, Torralbo A, Jolley KA, Kelly DJ, Bentley SD, Maiden MC, Parkhill J, Falush D: **Genome-wide association study identifies vitamin B5 biosynthesis as a host specificity factor in Campylobacter**. *Proc Natl Acad Sci USA* 2013, **110**:11923-11927.

11. Thompson O, Edgley M, Strasbourger P, Flibotte S, Ewing B, Adair R, Au V, Chaudhry I, Fernando L, Hutter H *et al.*: **The million mutation project: a new approach to genetics in *Caenorhabditis elegans***. *Genome Res* 2013, **23**:1749-1762.

12. Walker TM, Ip CL, Harrell RH, Evans JT, Kapatai G, Dedicoat MJ, Eyre DW, Wilson DJ, Hawkey PM, Crook DW *et al.*: **Whole-genome sequencing to delineate *Mycobacterium tuberculosis* outbreaks: a retrospective observational study**. *Lancet Infect Dis* 2013, **13**:137-146.

13. Casali N, Nikolayevskyy V, Balabanova Y, Harris SR, Ignatyeva O, Kontsevaya I, Corander J, Bryant J, Parkhill J, Nejentsev S *et al.*: **Evolution and transmission of drug-resistant tuberculosis in a Russian population**. *Nat Genet* 2014, **46**:279-286.

14. Chewapreecha C, Harris SR, Croucher NJ, Turner C, Marttinen P, Cheng L, Pessia A, Aanensen DM, Mather AE, Page AJ *et al.*: **Dense genomic sampling identifies highways of pneumococcal recombination**. *Nat Genet* 2014, **46**:305-309.

15. Nasser W, Beres SB, Olsen RJ, Dean MA, Rice KA, Long SW, Kristinsson KG, Gottfredsson M, Vuopio J, Raisanen K *et al.*: **Evolutionary pathway to increased virulence and epidemic group A Streptococcus disease derived from 3615 genome sequences**. *Proc Natl Acad Sci USA* 2014, **111**:E1768-E1776.

16. The biggest genome sequencing projects: the uber-list! http://pathogenomics.bham.ac.uk/blog/2013/12/the-biggest-genome-sequencing-projects-the-uber-list/.

17. 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**:6403-6405 discusion 6405.

18. •• Koren S, Harhay GP, Smith TP, Bono JL, Harhay DM, McVey SD, Radune D, Bergman NH, Phillippy AM: **Reducing assembly complexity of microbial genomes with single-molecule sequencing**. *Genome Biol* 2013, **14**:R101.

The 'assembly complexity' paper demonstrates that a majority of microbial genomes can be fully resolved to finished-grade accuracy using PacBio SMRT sequencing based on an anaylsis of all genomes in Genbank as of 2012. Genomes are classified into three complexity classes based on their repeat content to predict expected assembly quality. An updated version of PBcR for non-hybrid hierarchical assembly is also presented.

19. Genomes OnLine Database (GOLD). http://www.genomesonline.org/cgi-bin/GOLD/Search.cgi.

20. Miller JR, Koren S, Sutton G: **Assembly algorithms for next-generation sequencing data**. *Genomics* 2010, **95**:315-327.

21. Nagarajan N, Pop M: **Sequence assembly demystified**. *Nat Rev Genet* 2013, **14**:157-167.

22. Pop M: **Genome assembly reborn: recent computational challenges**. *Brief Bioinform* 2009, **10**:354-366.

23. Phillippy AM, Schatz MC, Pop M: **Genome assembly forensics: finding the elusive mis-assembly**. *Genome Biol* 2008, **9** R55.

24. Salzberg SL, Phillippy AM, Zimin A, Puiu D, Magoc T, Koren S, Treangen TJ, Schatz MC, Delcher AL, Roberts M: **GAGE: a critical evaluation of genome assemblies and assembly algorithms**. *Genome Res* 2012, **22**:557-567.

25. Magoc T, Pabinger S, Canzar S, Liu X, Su Q, Puiu D, Tallon LJ, Salzberg SL: **GAGE-B: an evaluation of genome assemblers for bacterial organisms**. *Bioinformatics* 2013, **29**:1718-1725.

26. Kingsford C, Schatz MC, Pop M: **Assembly complexity of prokaryotic genomes using short reads**. *BMC Bioinform* 2010, **11**:21.

27. Treangen TJ, Abraham A.-L.L., Touchon M, Rocha EP: **Genesis, effects and fates of repeats in prokaryotic genomes**. *FEMS Microbiol Rev* 2009, **33**:539-571.

28. van Belkum A, Scherer S, van Alphen L, Verbrugh H: **Short-sequence DNA repeats in prokaryotic genomes**. *Microbiol Mol Biol Rev* 1998, **62**:275-293.

29. •• Chin CS, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, Clum A, Copeland A, Huddleston J, Eichler EE *et al.*: **Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data**. *Nat Methods* 2013, **10**:563-569.

The 'HGAP' paper demonstrates that a non-hybrid approach is possible for assembling PacBio SMRT sequences. Major advances presented are a new consensus method and the Quiver algorithm, which uses a statistical model of SMRT sequencing to generate high-accuracy base calls.

30. •• Ribeiro FJ, Przybylski D, Yin S, Sharpe T, Gnerre S, Abouelleil A, Berlin AM, Montmayeur A, Shea TP, Walker BJ *et al.*: **Finished bacterial genomes from shotgun sequence data**. *Genome Res* 2012, **22**:2270-2277.

The 'ALLPATHS' paper presents the first recipe for complete assemblies of prokaryotic genomes. It relies on a combination of short reads, paired-ends, and long reads to demonstrate finished-level accuracy on several genomes.

31. Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, Peluso P, Rank D, Baybayan P, Bettman B *et al.*: **Real-time DNA sequencing from single polymerase molecules**. *Science* 2009, **323**:133-138.

32. • Chin CS, Sorenson J, Harris JB, Robins WP, Charles RC, Jean-Charles RR, Bullard J, Webster DR, Kasarskis A, Peluso P *et al.*: **The origin of the Haitian cholera outbreak strain**. *N Engl J Med* 2011, **364**:33-42.

The *V. cholerae* paper is the first large-scale study using PacBio SMRT sequencing. The analysis was limited to reference-based mapping as the sequences were relatively short and noisy.

33. Ono Y, Asai K, Hamada M: **PBSIM: PacBio reads simulator — toward accurate genome assembly**. *Bioinformatics* 2013, **29**:119-121.

34. •• Koren S, Schatz MC, Walenz BP, Martin J, Howard JT, Ganapathy G, Wang Z, Rasko DA, McCombie WR, Jarvis ED, Phillippy AM: **Hybrid error correction and de novo assembly of single-molecule sequencing reads**. *Nat Biotechnol* 2012, **30**:693-700.

The 'PBcR' paper presents the first hierachical assembly method and demonstrates that long, noisy reads can be used for assembly after correction. Low-coverage, long reads are shown to significantly improve assembly quality versus short reads alone. It also demonstrates the use of PacBio SMRT sequencing for assembling the parrot genome and RNA-Seq analysis of the corn transcriptome.

35. •• English AC, Richards S, Han Y, Wang M, Vee V, Qu J, Qin X, Muzny DM, Reid JG, Worley KC, Gibbs RA: **Mind the gap: upgrading genomes with Pacific biosciences RS long-read sequencing technology**. *PLoS One* 2012, **7**:e84776.

The 'PBJelly' paper presents the first assembly-boosting method for long read sequencing data, the paper demonstrates over 50% gap closure on eukaryotic model organisms.

36. •• Berlin K, Koren S, Chin C-S, Drake J, Landolin JM, Phillippy AM: **Assembling large genomes with single-molecule sequencing and locality sensitive hashing**. *bioRxiv* 2014.

The 'MHAP' paper presents an efficient method for overlapping long, noisy sequences, making non-hybrid assembly of large genomes computationally tractable. Assemblies of several model organisms are presented, including a haploid human cell line.

37. •• Lee H, Gurtowski J, Yoo S, Marcus S, McCombie WR, Schatz M: **Error correction and assembly complexity of single molecule sequencing reads**. *bioRxiv* 2014.

The 'ECTools' paper includes a study of assembly complexity for eukaryotic genomes and predicts assembly quality for novel genomes based on genome size. The study introduces a metric for evaluation of assembly performance (defined as N50/chromosome N50) and extends the hierachical hybrid approach to use pre-assembled sequence data.

38. Ross MG, Russ C, Costello M, Hollinger A, Lennon NJ, Hegarty R, Nusbaum C, Jaffe DB: **Characterizing and measuring bias in sequence data**. *Genome Biol* 2013, **14**:R51.

39. Voskoboynik A, Neff NF, Sahoo D, Newman AM, Pushkarev D, Koh W, Passarelli B, Fan HC, Mantalas GL, Palmeri KJ *et al.*: **The genome sequence of the colonial chordate, *Botryllus schlosseri***. *Elife* 2013, **2**:e00569.

40. McCoy RC, Taylor RW, Blauwkamp TA, Kelley JL, Kertesz M, Pushkarev D, Petrov DA, Fiston-Lavier A-S: **Illumina TruSeq synthetic long-reads empower de novo assembly and resolve complex, highly repetitive transposable elements**. *bioRxiv* 2014.

41. TruSeq Synthetic Long-Read DNA Library Prep Kit. http://www.illumina.com/products/truseq-synthetic-long-read-kit.ilmn.

42. Schneider GF, Dekker C: **DNA sequencing with nanopores**. *Nat Biotechnol* 2012, **30**:326-328.

43. The second Oxford Nanopore read ever, published, figshare. http://dx.doi.org/10.6084/m9.figshare.1881060.

44. A *P. aeruginosa* serotype-defining single read from our first Oxford Nanopore run. http://dx.doi.org/10.6084/m9.figshare.1052996.

45. Quick J, Quinlan A, Loman N: **A reference bacterial genome dataset generated on the MinION(TM) portable single-molecule nanopore sequencer**. *GigaScience* 2014, **3**:22.

46. • Boetzer M, Pirovano W: **SSPACE-LongRead: scaffolding bacterial draft genomes using long read sequence information**. *BMC Bioinform* 2014, **15**:211.

The 'SSPACE-LongRead' paper extends long read support for SSPACE, a general-purpose scaffolder. It demonstrates that low coverage data can improve assembly continuity but that deeper coverage is needed to resolve all repeats.

47. • Rasko DA, Webster DR, Sahl JW, Bashir A, Boisen N, Scheutz F, Paxinos EE, Sebra R, Chin CS, Iliopoulos D *et al.*: **Origins of the *E. coli* strain causing an outbreak of hemolytic-uremic syndrome in Germany**. *N Engl J Med* 2011, **365**:709-717.

The *E. coli* paper is the first large-scale study of PacBio SMRT sequencing that includes bacterial assembly. The study utilizes the hierachical hybrid approach to generate a reference assembly, which was combined with mapping approaches to study variation.

48. Myers EW: **A whole-genome assembly of *Drosophila***. *Science* 2000, **287**:2196-2204.

49. • Au KF, Underwood JG, Lee L, Wong WH: **Improving PacBio long read accuracy by short read alignment**. *PLoS One* 2012, **7**:e46679.

The 'LSC' paper presents a hierachical method focused on RNA-seq data as opposed to assembly. Homopolymers are compressed in the sequences to accelerate correction.

50. • Salmela L, Rivals E: **LoRDEC: accurate and efficient long read error correction**. *Bioinformatics* 2014.

The approach combining read-threading with hierachical hybrid algorithms. A de Bruijn graph is used to generate corrected sequences for long reads. At the time of writing, no assembly results have been presented.

51. • Hackl T, Hedrich R, Schultz J, Förster F: **proovread: large-scale high-accuracy PacBio correction through iterative short read consensus**. *Bioinformatics* 2014, **30**:3004-3011.

A hierachical hybrid approach utilizing an iterative mapping strategy at decreasing stringency to achieve its speedup. No assembly results were presented in the manuscript.

52. • Ye C, Hill C, Koren S, Ruan J, (Sam)Ma Z, Yorke JA, Zimin A: **DBG2OLC: Efficient assembly of large genomes using the compressed overlap graph**. 2014, arXiv:14102801.

A hybrid approach utilizing short-read contigs to identify long-read overlaps and build an overlap graph. The publication presents several eukaryotic assemblies, including a human cell line, which required only a few hours of runtime on a single server (assuming the Illumina data was already assembled).

53. Chevreux B, Pfisterer T, Drescher B, Driesel AJ, Muller WE, Wetter T, Suhai S: **Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs**. *Genome Res* 2004, **14**:1147-1159.

54. Myers G: **A de novo whole genome shotgun assembler for noisy long read data**. *AGBT 2014* 2014.

55. •• Myers G: **Efficient local alignment discovery amongst noisy long reads**. In *Algorithms in Bioinformatics.* Edited by Brown D, Morgenstern B.. **Lecture Notes in Computer Science**. Berlin, Heidelberg: Springer; 2014:52-67.

The 'DALIGNER' paper describes an efficient method to identify overlaps and alignments between long, noisy sequences. While not currently used by an assembler, it promises to make non-hybrid correction tractable for large genomes.

56. • Miyamoto M, Motooka D, Gotoh K, Imai T, Yoshitake K, Goto N, Iida T, Yasunaga T, Horii T, Arakawa K *et al.*: **Performance comparison of second- and third-generation sequencers using a bacterial genome with two chromosomes**. *BMC Genom* 2014, **15**:699.

The 'Sprai' paper describes a non-hybrid assembly approach. Like HGAP, ECTools, and PBcR, it relies on Celera Assembler for assembly of corrected data. It utilizes BLAST for alignments rather than BLASR. The paper presents a finished bacterial genome that HGAP had previously been unable to resolve.

57. Chaisson MJ, Tesler G: **Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory**. *BMC Bioinform* 2012, **13**:238.

58. PBcR/MHAP User Guide. http://wgs-assembler.sourceforge.net/wiki/index.php?title=PBcR.

59. •• Bashir A, Klammer AA, Robins WP, Chin CS, Webster D, Paxinos E, Hsu D, Ashby M, Wang S, Peluso P *et al.*: **A hybrid approach for the automated finishing of bacterial genomes**. *Nat Biotechnol* 2012, **30**:701-707.

The 'AHA' paper presents the first scaffolding approach for long reads and presents a finished bacterial genome. While not automated, the combination of short and long sequence data was sufficient to resolve the genome.

60. Pop M, Kosack DS, Salzberg SL: **Hierarchical scaffolding with Bambus**. *Genome Res* 2004, **14**:149-159.

61. Powers JG, Weigman VJ, Shu J, Pufky JM, Cox D, Hurban P: **Efficient and accurate whole genome assembly and methylome profiling of *E. coli***. *BMC Genom* 2013, **14**:675.

62. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD *et al.*: **SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing**. *J Comput Biol* 2012, **19**:455-477.

63. • Deshpande V, Fung EK, Pham S, Bafna V: **Cerulean: a hybrid assembly using high throughput short and long reads**. In *Algorithms in Bioinformatics.* Edited by Darling A, Stoye J.. **Lecture Notes in Computer Science**. Berlin, Heidelberg: Springer; 2013:349-363.

The 'Cerulean' paper describes a read threading approach that uses long reads to resolve an assembly graph. It is the first read threading approach to not rely on paired-end information to resolve repeats; however, no finished genomes were demonstrated.

64. Pevzner PA, Tang H, Waterman MS: **An Eulerian path approach to DNA fragment assembly**. *Proc Natl Acad Sci USA* 2001, **98**:9748-9753.

65. Nagarajan N, Pop M: **Parametric complexity of sequence assembly: theory and applications to next generation sequencing**. *J Comput Biol* 2009, **16**:897-908.

66. Satou K, Shiroma A, Teruya K, Shimoji M, Nakano K, Juan A, Tamotsu H, Terabayashi Y, Aoyama M, Teruya M *et al.*: **Complete genome sequences of eight helicobacter pylori strains with different virulence factor genotypes and methylation profiles, isolated from patients with diverse gastrointestinal diseases on Okinawa Island, Japan determined using PacBio single-molecule real-time technology**. *Genome Announc* 2014, **2**.

67. Harhay GP, McVey DS, Koren S, Phillippy AM, Bono J, Harhay DM, Clawson ML, Heaton MP, Chitko-McKown CG, Korlach J, Smith TP: **Complete closed genome sequences of three bibersteinia trehalosi nasopharyngeal isolates from cattle with shipping fever**. *Genome Announc* 2014, **2**.

68. Harhay GP, Murray RW, Lubbers B, Griffin D, Koren S, Phillippy AM, Harhay DM, Bono J, Clawson ML, Heaton MP *et al.*: **Complete closed genome sequences of four mannheimia varigena isolates from cattle with shipping fever**. *Genome Announc* 2014, **2**.

69. Brown SD, Nagaraju S, Utturkar S, De Tissera S, Segovia S, Mitchell W, Land ML, Dassanayake A, Kopke M: **Comparison of single-molecule sequencing and hybrid approaches for finishing the genome of *Clostridium autoethanogenum* and analysis of CRISPR systems in industrial relevant Clostridia**. *Biotechnol Biofuels* 2014, **7**:40.

70. Harhay GP, Koren S, Phillippy AM, McVey DS, Kuszak J, Clawson ML, Harhay DM, Heaton MP, Chitko-McKown CG, Smith TPL: **Complete Closed genome sequences of mannheimia haemolytica serotypes A1 and A6, isolated from cattle**. *Genome Announc* 2013, **1**.

71. PBcR/MHAP Whole Genome Assembly AWS Image. https://console.aws.amazon.com/ec2/v2/home?region=us-east-1#Images:filter=all-images;platform=all-platforms;visibility=public-images;search=ami-234e514a.

72. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J, Young SK, Earl AM: **Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement**. *PLoS One* 2014.

73. Teague B, Waterman MS, Goldstein S, Potamousis K, Zhou S, Reslewic S, Sarkar D, Valouev A, Churas C, Kidd JM *et al.*: **High-resolution human genome structure by single-molecule analysis**. *Proc Natl Acad Sci USA* 2010, **107**:10848-10853.

74. Conlan S, Thomas PJ, Deming C, Park M, Lau AF, Dekker JP, Snitkin ES, Clark TA, Luong K, Song Y *et al.*: **Single-molecule sequencing to track plasmid diversity of hospital-associated carbapenemase-producing Enterobacteriaceae**. *Sci Transl Med* 2014, **6**:254ra126.

75. Burton JN, Adey A, Patwardhan RP, Qiu R, Kitzman JO, Shendure J: **Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions**. *Nat Biotechnol* 2013, **31**:1119-1125.

76. Kaplan N, Dekker J: **High-throughput genome scaffolding from in vivo DNA interaction frequency**. *Nat Biotechnol* 2013, **31**:1143-1147.

77. Selvaraj SJRD, Bansal V, Ren B: **Whole-genome haplotype reconstruction using proximity-ligation and shotgun sequencing**. *Nat Biotechnol* 2013, **31**:1111-1118.

78. 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.

79. Treangen TJ, Sommer DD, Angly FE, Koren S, Pop M: **Next generation sequence assembly with AMOS**. *Curr Protocols Bioinform* 2011:18. Chapter 11:Unit 11.8.

80. Schatz MC, Phillippy AM, Sommer DD, Delcher AL, Puiu D, Narzisi G, Salzberg SL, Pop M: **Hawkeye and AMOS: visualizing and assessing the quality of genome assemblies**. *Brief Bioinform* 2011, **14**:213-224.

81. Koren S, Treangen TJ, Pop M: **Bambus 2: scaffolding metagenomes**. *Bioinformatics* 2011, **27**:2964-2971.

82. Nijkamp JF, Pop M, Reinders MJT, de Ridder D: **Exploring variation-aware contig graphs for (comparative) metagenomics using MaryGold**. *Bioinformatics* 2013, **29**:2826-2834.

83. Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJM, Birol I: **ABySS: a parallel assembler for short read sequence data**. *Genome Res* 2009, **19**:1117-1123.

84. Sanger F, Nicklen S, Coulson AR: **DNA sequencing with chain-terminating inhibitors**. *Proc Natl Acad Sci USA* 1977.

85. First Data Set from FastTrack Long Reads Early Access Service. http://blog.basespace.illumina.com/2013/07/22/first-data-set-from-fasttrack-long-reads-early-access-service/.

86. Jaffe D: **Assembly of bacterial genomes using long nanopore reads**. *AGBT 2014* 2014.

87. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool**. *J Mol Biol* 1990, **215**:403-410.

88. Kielbasa SM, Wan R, Sato K, Horton P, Frith MC: **Adaptive seeds tame genomic sequence comparison**. *Genome Res* 2011, **21**:487-493.