Annual Review of Animal Biosciences

New Approaches for Genome Assembly and Scaffolding

Edward S. Rice[1] and Richard E. Green[1,2]

[1]Department of Biomolecular Engineering, University of California, Santa Cruz, California 95064, USA; email: esrice@soe.ucsc.edu, ed@soe.ucsc.edu

[2]Dovetail Genomics, LLC, Santa Cruz, California 95060, USA

Annu. Rev. Anim. Biosci. 2019. 7:21.1–21.24

The *Annual Review of Animal Biosciences* is online at animal.annualreviews.org

https://doi.org/10.1146/annurev-animal-020518-115344

Keywords

genome, assembly, scaffolding, high-throughput sequencing, comparative genomics

Abstract

Affordable, high-throughput DNA sequencing has accelerated the pace of genome assembly over the past decade. Genome assemblies from high-throughput, short-read sequencing, however, are often not as contiguous as the first generation of genome assemblies. Whereas early genome assembly projects were often aided by clone maps or other mapping data, many current assembly projects forego these scaffolding data and only assemble genomes into smaller segments. Recently, new technologies have been invented that allow chromosome-scale assembly at a lower cost and faster speed than traditional methods. Here, we give an overview of the problem of chromosome-scale assembly and traditional methods for tackling this problem. We then review new technologies for chromosome-scale assembly and recent genome projects that used these technologies to create highly contiguous genome assemblies at low cost.

INTRODUCTION

The first projects to sequence and assemble the genomes of multicellular eukaryotes, starting with fruit fly in 2000 (1), human in 2001 (2), and mouse in 2002 (3), used capillary sequencing (also known as Sanger sequencing) (4) to read the sequence of many short, cloned DNA fragments. With automated Sanger sequencing, reading one million bases of DNA (1 Mb) cost approximately $1,500 and took more than a day when highly parallelized. Thus, reading enough copies of the 3 billion–base (3 Gb) human genome to accurately assemble it cost billions of dollars and took years of machine time. Genome projects were therefore the domain of a few large institutions and focused on model organisms commonly used in biological research.

New sequencing technologies, first large-scale pyrosequencing (5) and later SOLiD (6), Ion Torrent (7), and Solexa sequencing (8), brought down the cost and time required to generate genome-scale sequencing data. These technologies put genome sequencing within the reach of smaller labs studying nonmodel organisms. In 2007, pyrosequencing performed on the 454 high-throughput sequencer (5) was used to sequence a human genome to 7.4× coverage in 2 months, with one-tenth the cost of Sanger sequencing (9). By 2010, the Illumina HiSeq 2000 could sequence DNA more than 10,000 times faster than automated Sanger sequencing at less than 1/10,000 of the cost (10–13).

As a result of these new technologies, the number of published vertebrate genomes has increased greatly in the past decade (**Figure 1*a***), enabling genomic approaches to address questions in many research domains. For example, complete genomes have allowed study of the deep history of fast-evolving viruses, which sometimes integrate into host genomes, creating a time capsule of their organization since the integration event. This comparative genomics approach revealed that endogenous hepatitis B has been part of reptilian genomes for more than 200 million years (14). Genome assemblies are used as alignment references for sequences from different populations (resequencing projects) or related species, allowing discoveries such as the history, timing, and location of admixture events, including those between brown and polar bears (15), humans and Neanderthals (16–18), multiple species of Darwin's finches (19), and two species of monkey-flower (20).

Complete genomes have revolutionized the practical application of molecular biology and genetics research. Genome sequencing combined with the powerful CRISPR/Cas9 editing approach (21, 22) allows the function of any specific gene to be assayed by making targeted changes. This approach, coupled with complete genome sequence, could simplify and accelerate experimental analysis of gene function (23, 24).

Although technological advances have made sequencing DNA much cheaper and faster, short-read, high-throughput sequencing exacerbates the central challenge in genome assembly: accurate assembly of genomes that are often highly repetitive (**Figure 2**). Consequently, the contiguity of new genome assemblies decreased as high-throughput sequencing was widely adopted (**Figure 1*b*,*c***) (25–27), despite the importance of highly contiguous genomes for many comparative genomics analyses (28). One cause of this reduction in contiguity is the shift away from the way DNA was prepared for Sanger sequencing: cloning DNA into plasmid libraries. Plasmid libraries enabled generation of mate-pair data, i.e., generating sequencing reads from both ends of the plasmid insert, for little additional cost relative to single-end sequencing. Many early genome assemblies benefited from mate-pair data whose insert sizes were several kilobases long, as a by-product of the necessity of bacterial cloning for DNA amplification. Another factor in the recent reduction in assembly contiguity is economical. Previously, the input Sanger sequence data for genome assembly were so expensive that generating additional scaffolding data to improve contiguity did not substantially alter the total cost of a genome assembly project.

Whereas generating DNA sequence data became fast, easy, and economical, approaches for generating scaffolding data necessary for chromosome-scale genome assembly remained time

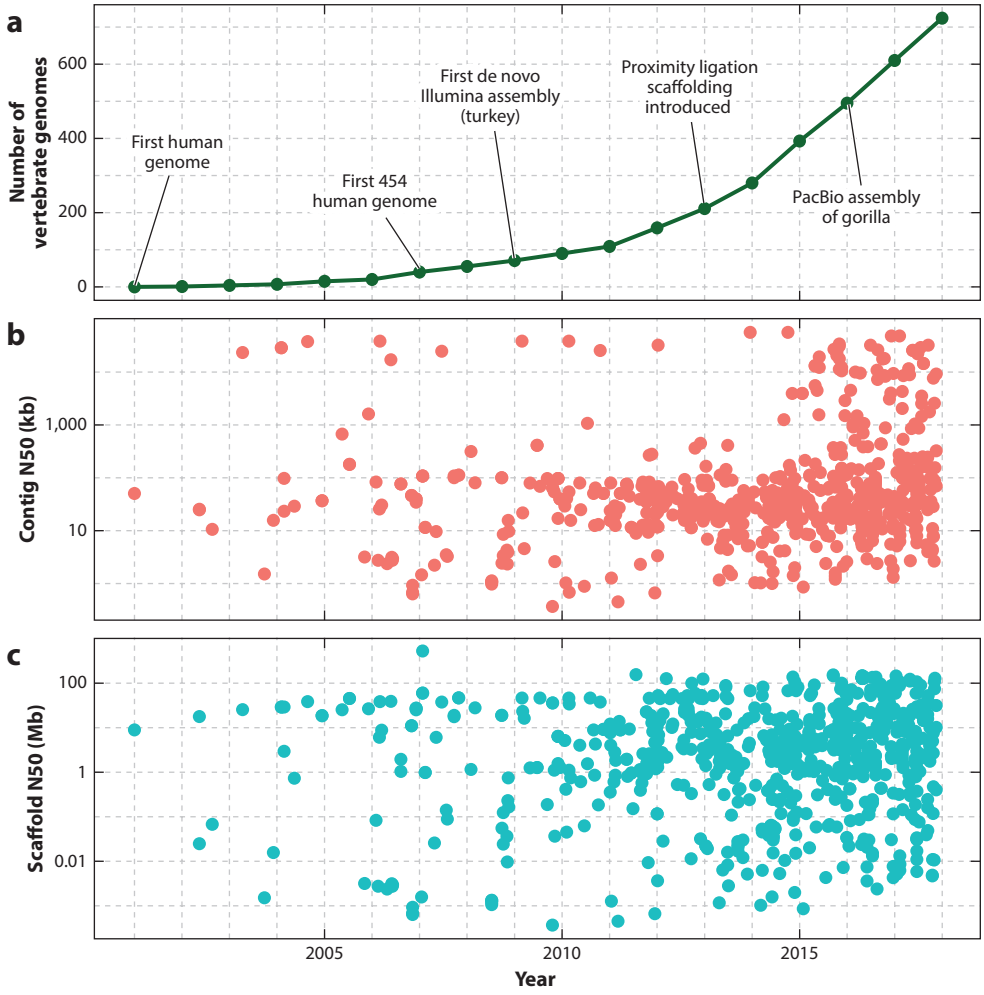

Figure 1

Timeline and statistics of vertebrate genome assemblies deposited in the National Center for Biotechnology Information's Genbank. Although second-generation sequencing has allowed more genomes to be published each year by making sequencing faster and cheaper, it has not increased the contiguity of published genomes. (*a*) Number of vertebrate genome assemblies available on Genbank at the end of each year, showing accelerating growth over the past decade. (*b*) Contig and (*c*) scaffold N50s of all vertebrate genome assemblies deposited in Genbank per year.

consuming, labor intensive, and expensive. Thus, many genome projects chose to forego scaffolding or map data and consequently produced highly fragmented draft genomes. Recently, several new technologies have been developed that produce data that can be used to increase the contiguity and accuracy of genomes with creative uses of high-throughput or next-generation sequencing data.

Genome Contig Assembly

No technology currently exists that can read DNA from one end to the other of even moderately sized chromosomes, which are typically tens or hundreds of millions of base pairs long. All current

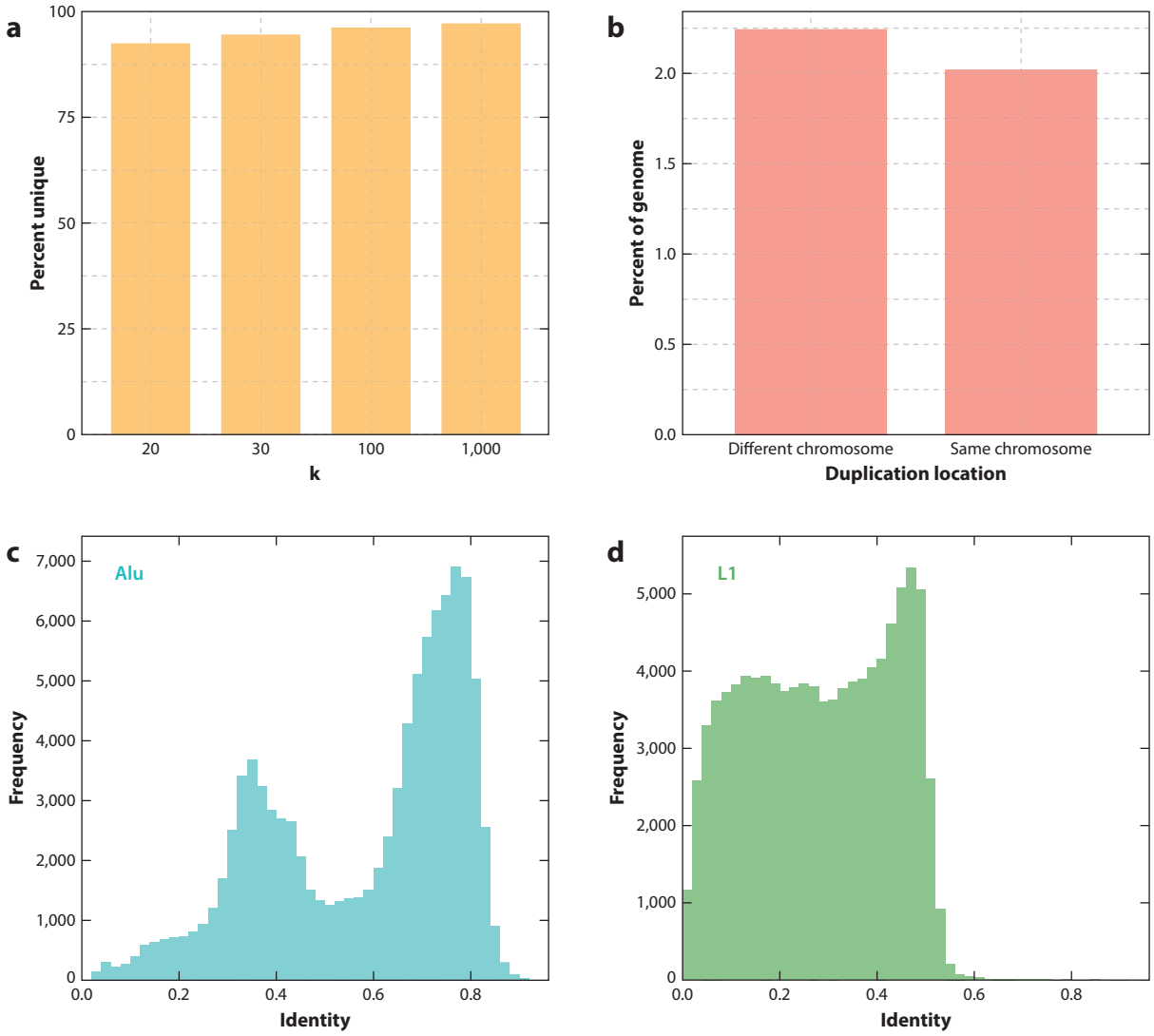

Figure 2

Repetitive content creates a challenge in genome assembly, as illustrated by the repetitive content of the human genome. (*a*) Percentage of k-mers in the human genome that occur only once, for different values of k. Even at k = 1,000, some k-mers appear multiple times in the genome. (*b*) Percentage of genome consisting of segmental duplications using alignments of 5-kb sequences with identity greater than 95%. (*c–d*) Distribution of alignment identities for 100,000 randomly sampled pairs of (*c*) Alu and (*d*) L1 repetitive elements.

approaches for genome assembly read many segments that are considerably shorter than chromosomes: hundreds of base pairs for Illumina (29), thousands or tens of thousands for PacBio (30, 31), and occasionally hundreds of thousands on the quickly evolving Oxford Nanopore (32) platform.

The process of converting input genomic DNA into sequencing libraries is necessarily platform dependent (**Figure 3**). However, in each case this involves ligation of adapter sequences to input genomic DNA. For the Illumina platform, these adapters contain polymerase chain reaction (PCR) primer sites used for in situ PCR amplification on a flow cell and sequencing primer

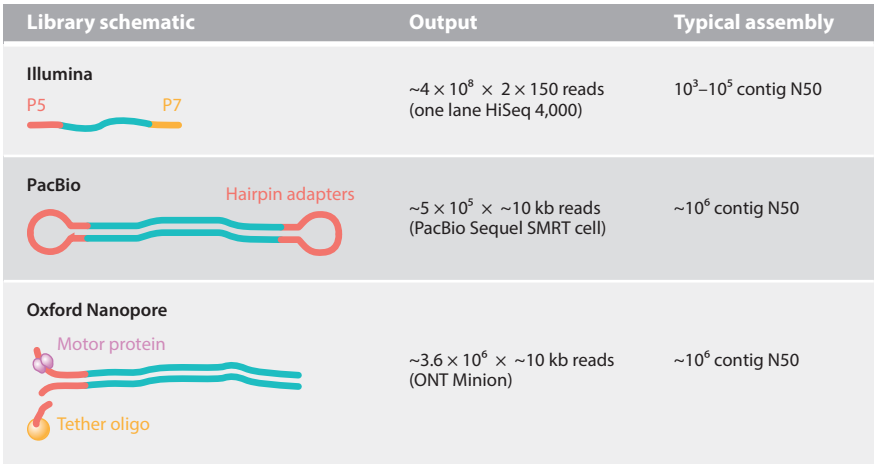

Figure 3

Overview of sequencing library architecture, output, and assembly results from three high-throughput sequencing technologies. For each sequencing platform, the data output column reflects the number and length of reads generated by one typical unit of sequencing. The typical contig N50 column summarizes typical results from de novo assembly projects using data only from the indicated platform, for example, Illumina (175), PacBio (176), and Oxford Nanopore (177).

sites for the sequencing by synthesis that follows. There are many creative approaches to generation of Illumina libraries that are designed to limit biases in DNA fragmentation and PCR (33–35). The Pacific Biosciences and Oxford Nanopore Technologies platforms are both long-read, single-molecule sequencers. For these platforms, high-quality library generation requires recovery of clean, high–molecular weight DNA (31, 36, 37).

The past decade has seen tremendous growth in the development of computational algorithms for generating sequences of contiguous segments of the genome (contigs) from these data (**Table 1**). Most first-generation assemblers were based on the overlap-layout-consensus approach (38), wherein input DNA sequence reads are compared, all versus all, in the overlap step. Thus, the time required for assembly via overlap-layout-consensus grows quadratically with the size of the input data. This approach is tractable for assembly of smaller numbers of long reads. It became intractable for the billions of input reads that are typically generated on the Illumina platform for genome assembly.

To address this limitation, several groups have written software that uses high-throughput Illumina sequence data to populate de Bruijn graphs or other graph structures (39–41). Typically, short words (k-mers) that are observed in the reads are the nodes of the graph, and edges are added when these k-mers are adjacent in sequence reads. In this process, each read is used to populate the graph but not compared directly to all the other reads. Thus, the algorithmic complexity of these graph-based assembly algorithms scales linearly (not quadratically) with the number of input DNA sequence reads. Importantly, because the nodes in these graphs are k-mers, sequence accuracy is important. A single base sequencing error can induce k false k-mers in the graph and the concomitant loss of k correct k-mers. The crux of these approaches is that overlapping reads are identified by virtue of containing some set of identical k-mers in identical order, but not by directly comparing the reads themselves.

Contig: a contiguous DNA sequence assembled from shorter reads based on overlaps between them

Whether by overlap-layout-consensus or graph-based methods, assembly proceeds by determining some number of contigs. The algorithms used for this step vary widely (**Table 1**), and the

Table 1 Commonly used assembly software

Software	URL and reference	Description
Short-read assembly software		
Velvet	**http://github.com/dzerbino/velvet** (168)	Original de Bruijn graph assembler
SOAPdenovo	**http://soap.genomics.org.cn/** (169)	De Bruijn graph assembler with error-correction step
Meraculous	**https://jgi.doe.gov/data-and-tools/meraculous/** (170)	Hybrid k-mer/read-based
ALLPATHS-LG	**http://software.broadinstitute.org/allpaths-lg/blog/** (171)	Uses unipath graph to collapse repeats
SGA	**https://github.com/jts/sga** (172)	Uses string graphs
ABySS	**https://github.com/bcgsc/abyss** (173)	Represents de Bruijn graph with a Bloom filter
DISCOVAR de novo	**https://software.broadinstitute.org/software/discovar/blog/** (174)	Requires 250-bp PCR-free reads
Supernova	**https://github.com/10XGenomics/supernova** (149)	Assembles 10× linked reads
Long-read assembly software		
HGAP	**https://github.com/PacificBiosciences/Bioinformatics-Training/wiki/HGAP** (124)	Error correction, overlap-layout-consensus assembly, and polishing workflow
Canu	**https://github.com/marbl/canu** (125)	K-mer-based overlap computation
FALCON	**https://github.com/PacificBiosciences/FALCON** (103)	Assembles phased diploid genomes
Flye	**https://github.com/fenderglass/Flye** (129)	Uses A-Bruijn graph
Miniasm	**https://github.com/lh3/miniasm** (128)	Fast, but no error correction
Polishing software		
Pilon	**https://github.com/broadinstitute/pilon** (133)	Uses short-read alignments to correct errors
Arrow	**https://github.com/PacificBiosciences/GenomicConsensus**	Hidden Markov model and long-read alignments
Nanopolish	**https://github.com/jts/nanopolish** (115)	Nanopore only; uses original voltage data to correct errors

optimal strategy depends on the genome to be assembled, as well as the data type and quality available as input. In practice, contig assembly generally produces thousands of contigs whose order and orientation relative to one another cannot be further described. In contrast to first-generation genome assembly, in which contigs generally ended where sequence coverage was too low to identify further overlapping reads, contig breaks from high-coverage short-read sequencing generally contain repetitive sequence. That is, contigs usually terminate not for lack of data representing those regions of the genome but rather because the regions themselves are too repetitive to determine how to extend them (42, 43).

Why Is Chromosome-Scale Assembly Important?

Two primary goals of many de novo genome assembly projects are to learn the sequence of all the genes in a genome and to have a reference genome sequence to which other individuals can be compared. Knowing the sequences of genes is useful for many purposes, such as comparing protein sequences between related species to learn how they have evolved or performing gene expression studies using RNA sequencing. Having a reference genome to compare other individuals to is

useful for learning about the population genetics of a species through calculation of statistics such as nucleotide diversity. Therefore, even a fragmented genome can be useful for many applications so long as it is contiguous enough to avoid splitting genes between scaffolds. However, many comparative genomics applications of reference genomes, such as studying chromosome-scale evolution and inferring ancestral karyotypes, require highly contiguous genome assemblies (44, 45).

Scaffold: a DNA sequence containing multiple contigs in the correct order and orientation, with gaps in between them; the process of assembling contigs into scaffolds is also called scaffolding

***Cis*-regulatory elements and the complexity of regulatory architecture.** Knowing the coding sequence of a gene may not provide information necessary for learning the conditions under which the gene is expressed. *Cis*-regulation of gene expression can be affected over large genomic distances, such as with interactions between enhancers and promoters, which can be more than 1 Mb apart on a chromosome (46, 47). These interactions are often able to take place owing to the physical organization of chromosomes bringing enhancers and promoters into close physical proximity (48).

The physical organization of chromosomes into domains of various sizes, and how this structure regulates gene expression, is currently an important area of inquiry, and long-range assemblies of nonmodel organisms have allowed important insights into this subject. For example, studying how chromatin architecture, and thus gene expression, can be disrupted in the human malaria parasite *Plasmodium falciparum* led to the development of several antimalarials (49), and examining estrogenic regulation of gene expression throughout long genomic regions in the American alligator gave insight into temperature-dependent sex determination (50). The chromosome-scale assembly and publication of the genomes of more organisms will allow future genomics projects to yield further insights into gene regulation across the tree of life.

Recombination. Because recombination during meiosis occurs on a chromosomal scale, a chromosome-scale assembly is necessary for studying recombination. Crossing over occurs at random, but not uniformly distributed, locations across the lengths of chromosomes (51, 52), with recombination occurring more frequently in hot spots (53). This nonuniform recombination landscape can lead to large differences in nucleotide diversity and effective population size across the length of a chromosome (51, 54, 55).

Genetic association studies. A genetic association study searches for genetic variants correlated with a trait. These studies are especially useful for traits that are multifactorial and polygenic, with no single variant being entirely predictive of the phenotype (56). Genetic association studies have been used to identify variants associated with susceptibility to many diseases in humans (57, 58), such as Parkinson's (59) and Crohn's (60). Owing to linkage disequilibrium, genetic association studies often find associations involving variants not related to the trait in question, but physically close to other variants that are causative of the trait in question (56). Interpreting such results is easier with a contiguous genome assembly because regions of linkage are less likely to be separated among different contigs. Genetic association studies have been performed in nonmodel organisms to identify variants associated with phenotypes, such as fire adaptation in lodgepole pines (61) and high-altitude adaptation in the ground tit (62).

Chromosome evolution. The organization of DNA into chromosomes changes over evolutionary time. Even closely related species often have different numbers of chromosomes. In some lineages, such as crocodilians, chromosomes evolve slowly (63, 64); in others, such as mammals, chromosomes evolve more quickly (65–67). Changes to chromosome structure are important in evolutionary biology because they can lead to reproductive isolation between species (68, 69) and accumulation of genetic differences between males and females of the same species (70). Studying

chromosome evolution requires chromosome-scale assemblies (reviewed in 28). Several studies have used assembled genomes to reconstruct ancestral karyotypes (44, 71–73) or to study the evolution of chromosomes more generally (45), and the quality of the reconstructions depends on the contiguity of the input assemblies.

TRADITIONAL APPROACHES FOR LONG-RANGE GENOME SCAFFOLDING

Genetic Mapping

Genetic markers that reside on the same chromosome are coinherited, except when separated by recombination (74). The chance that two markers will be recombined is a function of their genetic distance, which is correlated with physical distance. These central genetic facts provide a method for assigning contigs to linkage groups (which are often chromosomes) and for ordering contigs along chromosomes (75) that long predates the era of DNA sequencing.

Genetic maps can be used to assign contigs or scaffolds to chromosome locations by aligning the primer sequences of the genetic markers on the map to the assembly and then ordering and orienting the scaffolds based on the locations of these markers on the scaffolds, as recently reviewed by Mascher & Stein (76). These tasks are often performed with ad hoc scripts (77), but some software, such as Chromonomer (**http://catchenlab.life.illinois.edu/chromonomer/**), is available to automate the process.

Genetic maps have been used to assign scaffolds to chromosome locations during the assembly process for several genomes (77, 78), such as that of the horseshoe crab (79) and fugu (80). However, genetic mapping generally requires a large-scale genotyping effort. In organisms that have long generation times or are hard to raise, genetic mapping becomes intractable, or at least prohibitively expensive and time consuming. In addition, because recombination does not occur uniformly over the length of a chromosome but is more likely to occur in certain hot spots (51–53), the distances measured by a genetic map are not directly proportional to the base-pair distance between genes.

Radiation Hybrid Mapping

Radiation hybrid (RH) mapping is another method for discovering which genetic markers are in linked segments and for determining their order. Like genetic mapping, RH mapping estimates the distances between pairs of loci based on how often they are separated when their chromosome is broken. However, RH mapping uses radiation to break chromosomes instead of meiotic recombination. Cells containing the target genome are exposed to a lethal dose of radiation, which fragments their chromosomes. These fragments are then recovered in the cells of a different organism, which incorporate the fragments into their genomes with double-strand break–repair mechanisms. The hybrid cell lines are grown, and PCR is used to determine which markers are present in each cell line (81, 82). The distance between each pair of markers is then estimated based on the frequency with which that pair of markers appears together across all cell lines (83). These distances are then used to create a linkage map.

RH mapping was an improvement over genetic mapping because radiation fragments chromosomes in a more uniformly random fashion than meiotic recombination, and because genetic mapping requires genotyping many individuals while RH mapping does not require genotyping. However, this process is still not completely uniform (84). In addition, RH mapping requires culturing a large number of cell lines and testing for the presence of every marker in every cell line,

making it time consuming and expensive. Nonetheless, it can be more parallelizable and accurate than genetic mapping (85), and high-quality RH maps already exist for many species, so RH maps are still commonly used to assign scaffolds to chromosomes in genome assembly projects, such as for the most recent assemblies of the zebrafish (86), goat (87), chicken (88), and horse (89) genomes.

Fluorescence In Situ Hybridization Mapping

Fluorescence in situ hybridization (FISH) mapping uses fluorescently labeled probes to determine the locations of known markers along chromosomes. First, the DNA sequence of a marker is amplified using PCR and labeled with fluorescent dye to create probes. The probes are then hybridized with the target chromosomes through in situ complementary base pairing. The target chromosomes are karyotyped and viewed through a fluorescence microscope, which causes each probe to appear as a colored band on the chromosome with which it hybridized, giving the location of the marker in the genome (90). FISH mapping can be multiplexed by concurrently using different fluorescent dyes and partially automated using computer software (91, 92), although its parallelization is limited by the number of fluorescent dyes that can be used concurrently.

FISH mapping is a vital tool for assigning linkage groups from genetic or RH maps to chromosomes, as it actually places markers on specific chromosomes using a karyotype. However, the resolution of traditional FISH is approximately 1 Mb (93), making it less useful for determining the order of proximate markers. Modifications to the FISH protocol using less-condensed chromatin can increase the resolution to approximately 50 kb, but these methods cannot be used for chromosome assignment, as karyotyping requires condensed chromatin (94, 95). FISH is still used to assign scaffolds to chromosomes during genome assemblies, such as of the most recent tomato (96) and Asian seabass (97) genomes, and a cross-species form of FISH, zoo-FISH, has been used to validate assemblies (73).

Bacterial Artificial Chromosome–End Sequencing

Bacterial artificial chromosomes (BACs) were developed as a method for cloning large fragments of DNA (98) up to over 300 kb in length. BACs have been used extensively to guide genome assembly by BAC-end sequencing. In this approach, BAC clones are sequenced at both ends, using sequencing primers complementary to the BAC insertion site. This results in large-insert mate-pair data (99). These end sequences can be aligned to a contig assembly to order and orient contigs to form scaffolds (100).

BAC-end sequencing was used to assemble the first eukaryotic genomes, including *Drosophila* (1), human (2), and mouse (3), and existing BAC-end sequence libraries are still used in genome projects for quality-control purposes (50, 89). The disadvantages of BAC-end sequencing include the need for extensive cell culture work and the occasional presence of chimeric sequences in BAC libraries.

CREATING MORE CONTIGUOUS ASSEMBLIES WITH LONG READS

Perhaps the most obvious solution to genome assembly is to make the sequence reads themselves long enough to cover the sequences before, within, and after long repeats. These technologies are referred to collectively as long-read sequencing. The advantages of these methods are somewhat offset by their high error rates. In this section, we discuss the two current long-read sequencing technologies as well as software available for using these reads in assemblies.

PacBio Single-Molecule Real-Time Sequencing

N50: a statistic used to measure genome contiguity; if a set of sequences is ordered by length, N50 is the length for which the summed length of all sequences greater than or equal to that length is half of the length of the whole genome

Pacific Biosciences, Inc. published a new method for sequencing DNA in 2009 (30). This method, called single-molecule real-time (SMRT) sequencing, is distinguished by the lengths of its reads: A PacBio Sequel machine produces reads with N50 of approximately 15 kb (101), much longer than Illumina short reads (75–300 bp) or Sanger reads (~1 kb). These long reads are useful for assembly, as they are long enough to span many repetitive regions.

SMRT sequencing, like other methods such as Sanger and Illumina, uses a DNA polymerase to replicate the input DNA and fluorescently labeled dNTPs to determine the order in which bases are incorporated into the sequence (30). First, hairpin adapters are placed at both ends of each piece of DNA to create a circular molecule that can be sequenced several times. The redundant sequencing of template DNA is used to create a circular consensus sequence, thereby reducing sequencing errors. As the tethered polymerase moves the DNA being sequenced, each new base incorporation causes fluorescence to be localized to a sensor (102). In this way, PacBio sequencing observes the actions of the polymerase as the template moves through it.

Like other high-throughput sequencing technologies, the PacBio Sequel performs SMRT sequencing in a parallel fashion, with current versions of the machine containing one million Zero Mode Waveguides (sensor wells for fluorescence detection) per flow cell.

The length of SMRT sequencing reads is useful for assembly for several reasons. First, many classes of repetitive elements too long to be spanned by Illumina reads, such as DNA transposons and LINEs (43), are well within the ~15-kb read length N50 of SMRT-seq reads. Moreover, the presence of multi-kilobase genomic regions from the same haplotype in single reads can facilitate phased diploid assembly (103). Finally, long reads also allow the detection of large structural variations in the genome (104).

The primary limitations of SMRT compared with other technologies are reduced accuracy and increased cost per base pair compared with Illumina sequencing. The PacBio Sequel has an error rate of approximately 15%, with most of the errors being insertions and deletions, which are harder to detect and correct computationally than the base miscalls that characterize the error profile of Illumina short reads (101). This is an order of magnitude larger than the error rate of the Illumina HiSeq 2500, which is less than 1% (29). The cost of SMRT sequencing, at approximately $0.40/Mb, is also an order of magnitude higher than the cost of Illumina sequencing, at approximately $0.04/Mb (31).

Nanopore Sequencing

Unlike other sequencing technologies, nanopore sequencing does not rely on a DNA polymerase. Nanopore sequencing, commercialized by Oxford Nanopore Technologies as the MinION, GridION, and PromethION sequencers, instead reads the sequence of DNA by measuring voltage changes as a DNA strand moves through a pore embedded in a membrane (105). First, DNA is placed on one side of a membrane and broken into single strands. Voltage across the membrane causes the negatively charged DNA to move through the pore embedded in the membrane. When DNA is moving through the pore, this blocks ions in the solution from moving through the pore, which alters the current. By measuring these changes in current, the sequences of bases moving through the pore can be determined (106, 107). This allows long strands of DNA to be read, resulting in read lengths of 100 kb or longer, with the longest reported read over 2 Mb in length (108). Some library preparation techniques add a hairpin adapter to one end of each piece of DNA, allowing both strands to be read in one pass and producing redundancy to improve accuracy (32).

The length of nanopore reads gives them the same advantages in de novo assembly as SMRT-seq reads. However, while nanopore reads are generally longer than SMRT-seq reads, nanopore

error rates can also be higher than those of SMRT-seq (109), although error rates as low as 12% have been reported with new library preparation techniques and base calling algorithms (110). Owing to these errors, de novo assembly projects often require long nanopore reads to be corrected with short reads before (107, 111, 112) or after (113) assembly, although Giordano et al. (114) report that de novo assemblies using only nanopore reads are comparable in quality to those using only PacBio reads, especially when using new error-correction methods designed for the high error rate of nanopore sequencing (115, 116). Nanopore sequencing on the MinION has the additional advantage of a small initial investment—roughly $1,000 for a MinION—as well as portability, with a MinION being a pocket-sized USB device (114).

Algorithms and Software

The de Bruijn graph assembly framework is now commonly used for contig assembly, as it is well-suited to assembling the large number of highly accurate short reads produced by Illumina sequencers. The performance of de Bruijn graph assembly is dependent on read accuracy but not on read length, as reads are broken into shorter k-mers, and generally no allowance is made for sequencer errors when determining whether two k-mers overlap (117). However, this approach is not suitable for assembling long reads, as it neither handles their high error rate well nor leverages their length to increase contig size.

One solution to this problem is to use a hybrid assembly approach with both long and short reads. The accuracy of the short reads is used to decrease the error rate of the long reads from up to 20% to as low as 0.1%. Then, the corrected long reads are assembled using an algorithm such as overlap-layout-consensus. Koren et al. (118) implemented this approach in the software package PBcR. PBcR aligns high-accuracy short reads to low-accuracy long reads, using these alignments to determine a consensus sequence for each of the long reads. Then, assembly proceeds using the corrected long reads and the Celera Assembler (119), which was originally designed to assemble Sanger sequence. Another hybrid error-correction approach, ECTools (120), assembles the short reads into unitigs with Celera Assembler, aligns the long reads to the unitigs with MUMmer (121), and uses these alignments to correct the long reads. Both SPAdes (122) and dbg2olc (123) begin by assembling the short reads with a de Bruijn graph, and then SPAdes uses the long reads to scaffold the short-read assembly, whereas dbg2olc uses the long reads and the short-read assembly together to build an overlap graph.

It is also possible to assemble long reads without also using short reads. HGAP (124) divides the long reads into two sets based on size, aligns the shorter long reads to the longer long reads, uses consensus from these alignments to correct the longer long reads, and assembles the corrected reads with an overlap-layout-consensus assembler such as Celera Assembler.

Canu (125) is a fork of Celera Assembler designed specifically for low-identity long reads. It first uses the MHAP k-mer hashing algorithm (126) to compute overlaps between the error-prone input reads while attempting to avoid mistaken overlaps from repetitive regions and then uses these overlaps to correct the reads. Next, reads containing segments unsupported by overlaps are trimmed or broken into multiple reads. Finally, Canu uses a modified version of the best overlap graph algorithm (127) to assemble the corrected and trimmed reads into contigs.

Miniasm (128) assembles long reads without error correction by computing overlaps using a new mapping algorithm, minimap, which is designed to take into account the high error rates associated with long reads. Skipping the error-correction step, it is reported, allows miniasm to perform assemblies faster than other methods at the cost of creating assemblies with the same high error rates as the input reads. These errors in the assembly can later be corrected using other software. Flye (129) is an assembler that also skips the read-error-correction stage. However,

unlike any of the other nonhybrid approaches discussed here, Flye uses a de Bruijn graph–based algorithm rather than an overlap graph–based algorithm. The modified de Bruijn graph, called an A-Bruijn graph (130), has the repeat-resolving capabilities of a classic de Bruijn graph but is better able to handle read errors (129).

One of the advantages of long reads is that they contain the information necessary to resolve large structural variants between haplotypes. Some long-read assemblers attempt to produce assemblies representing both haplotypes in diploid genomes, especially where they are distinct. The FALCON assembler (103), for example, can produce a diploid assembly using a process modeled on HGAP (124): Reads are error corrected and then assembled into an overlap graph, but bubbles in the overlap graph are left intact. Then, heterozygous sites are marked, and the original reads are used to resolve the bubbles into multiple haplotypes.

Long reads can also be used to scaffold or fill gaps in existing assemblies. PBJelly (131) is a commonly used program that uses long reads to fill gaps. PBJelly maps long reads to an assembly and then uses gap-spanning reads to replace the Ns used to denote a gap between two ordered and oriented contigs with a better representation of the sequence in the gap. LINKS is a recently described approach for scaffolding genomes with long-read data (132). It uses a k-mer approach to describe long reads and contigs in terms of the k-mer content and the distance between k-mers. It then finds long reads with similar, but not necessarily matching, k-mer fingerprints that span contigs. This approach is less sensitive to sequencing errors than many k-mer methods, making LINKS useful for Oxford Nanopore Technologies long-read data.

Because of the high error rate of long reads, many sequencing errors end up in the assemblies produced by overlap-layout-consensus assemblers, even when input reads are error corrected before assembly. Therefore, a polishing step is often beneficial after assembly. Quiver and its successor Arrow, which are included in the HGAP package (124), are variant callers designed to use alignments of PacBio reads to an assembly, along with locations of known variants if available, to determine a consensus sequence. Pilon (133) can correct assembly errors using paired-end short-read alignments. For assemblies generated from nanopore reads, nanopolish (115) uses the original raw voltage data generated by the sequencer to recall bases in the context of the assembly.

NEW APPROACHES FOR LONG-RANGE GENOME SCAFFOLDING

Proximity Ligation

Inside a cell, the DNA in a chromosome must be physically folded and packed to fit into a small space. Parts of a chromosome that are far apart along the linear chromosome are often close together in physical space (134). Several methods have been invented for determining which parts of a chromosome are in close physical proximity, first including chromosome conformation capture in 2002 (135) and eventually including use of high-throughput sequencing to examine chromosome conformation over the entire genome in a method called Hi-C (136). The Hi-C protocol (**Figure 4*a***) generates an Illumina paired-end library wherein the reads in a pair represent genomic segments that were physically close. When mapped to a reference genome, Hi-C data can be used to determine the frequency of physical contact between any two regions of the genome.

Although these methods were invented to study how chromosomes fold, they can also be used for scaffolding an assembly. A key insight is that regions of the genome that are close together in sequence generally have more frequent physical contact than parts of the genome that are far apart in sequence. Nevertheless, regions of the same chromosome, even those megabases away, contact each other more often than they contact other chromosomes. These insights allow Hi-C data to be used to produce chromosome-scale scaffolds (137). For example, given Hi-C data and a set of scaffolds smaller than chromosomes, if two scaffolds have a high frequency of contacts

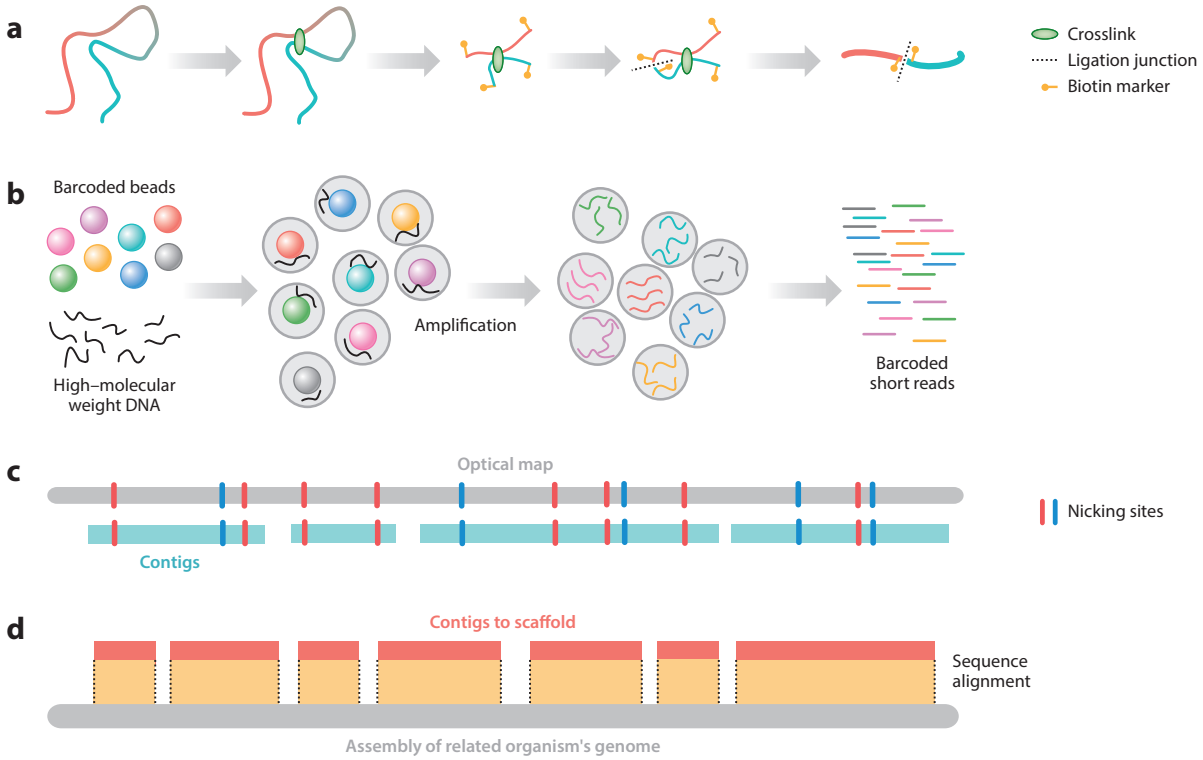

Figure 4

Overview of methods for long-range scaffolding. (*a*) In proximity ligation, chromatin is crosslinked and then restriction digested, ligated, and fragmented to create reads containing sequence from two different parts of the same chromosome. (*b*) In 10× linked-read sequencing, high–molecular weight DNA is combined with barcoded beads in oil droplets and then undergoes barcoding and amplification inside the droplets, resulting in reads with the same barcode that came from the same initial fragment of DNA. (*c*) BioNano optical maps are created by nicking high–molecular weight DNA with multiple nicking enzymes and attaching fluorescent markers at the nick sites. Contigs can then be aligned to the optical map by lining up nicking sequences in the contigs with the locations of fluorescent markers in the map. (*d*) In synteny-based approaches, contigs are mapped to the assembled genomes of one or more related species. These alignments imply the order and orientation of the aligned contigs.

between them, they are likely to come from the same chromosome. Additionally, the distribution of contacts across the lengths of scaffolds can be used to order and orient them into chromosome-length meta-scaffolds.

A confounding factor when using Hi-C data to scaffold a genome assembly is that the contact frequency between two loci on the same chromosome depends not only on the sequence distance between them but also on how the chromosome is folded. A modified proximity ligation technique called Chicago (138) uses reconstituted chromatin from purified high–molecular weight DNA rather than chromatin as it exists in the nuclei of cells. Because this chromatin is formed in vitro, it is not subject to the mechanisms that govern chromatin structure in living cells, so Chicago libraries have insert size distributions characterized by lower mean insert size and much smaller variance than those found in Hi-C libraries. This makes Chicago ideal for scaffolding of smaller input contigs.

Assembly algorithms. The first methods for scaffolding genomes using Hi-C reads were published in 2013 (137, 139, 140). LACHESIS (137) places input scaffolds into chromosome-length

meta-scaffolds via a three-step process: First, the input scaffolds are clustered into subsets such that scaffolds in the same subset share more Hi-C read pair links than they do with scaffolds in other subsets. Next, the scaffolds in each subset are ordered so that scaffolds with more links between them are more likely to be adjacent to each other. Finally, the positions where Hi-C reads map to each scaffold are used to determine the most likely orientation of each input scaffold on the output meta-scaffold. Phase Genomics Inc. has commercialized this approach and sells Hi-C scaffolding kits that include access to their proprietary scaffolding software, Proxima.

HiRise (138) uses a Chicago and/or Hi-C library to scaffold an input genome. HiRise aligns the proximity ligation reads to the input scaffolds and then estimates parameters for the distribution of insert sizes using pairs where both reads align to the same input scaffold. Next, a graph is created in which each vertex corresponds to an input scaffold and edges between nodes contain information about the alignment positions of read pairs that link the two scaffolds. HiRise clusters the graph into connected components representing output scaffolds by removing edges with low support. These clusters are then ordered by further pruning edges that connect nodes with high degree, as these represent loci that interact with each other too frequently to be explained by adjacency. Finally, the scaffolds are oriented using a dynamic program that maximizes the sum of the probabilities of the resulting insert sizes.

Another software package for scaffolding a genome with Hi-C reads is 3D-DNA (141). This pipeline uses a three-step approach in which Hi-C data are used first to identify and break misjoins in the input scaffolds, then to perform scaffolding, and last to collapse heterozygous regions into single haplotypes. In the misjoin-detection step, 3D-DNA breaks input scaffolds between any two loci with contact frequency below an estimated lower bound. In the scaffolding step, the broken input scaffolds are represented as a graph with edges weighted based on contact density between the half-scaffolds they connect, normalized by the incident edge with the maximum contact density. The graph is then traversed for maximum total edge weight to determine the order and orientation of input scaffolds. In the collapsing step, a combination of sequence similarity and Hi-C data are used to find and collapse uncollapsed heterozygous regions represented as different scaffolds into single haplotypes. In follow-up work (142), the authors performed de novo chromosome-scale assemblies of mammalian genomes using only 300 million shotgun reads and 100 million Hi-C reads, which cost less than $1,000.

SALSA (143) and SALSA2 (144) are Hi-C scaffolders that can take advantage of output from long-read contig assemblers and can use Chicago as well as Hi-C libraries. SALSA2 can take as input an assembly graph from a contig assembler rather than just the output contigs, giving it more complete information it can use to make scaffolding decisions. The disadvantage of this approach is that it requires the input contigs to be constructed using an overlap graph assembly of long reads, which are more expensive to produce than short reads. However, the authors report that for its intended use of scaffolding long-read assemblies, SALSA2 outperforms the other current state-of-the-art open-source Hi-C scaffolder 3D-DNA (141), with a large reduction in misjoins as well as order and orientation errors.

Linked-Read Sequencing

Linked-read sequencing is a method for generating short-read sequencing libraries in which multiple reads are barcoded to denote that they came from the same region of the genome. The first linked-read technology, CPT-seq, is a medium-range contiguity method designed for haplotyping (145) but also used for scaffolding (146). The transposase Tn5 is used to insert adapters to DNA, as in several library preparation protocols. But, unlike these protocols, the Tn5 is left bound to the DNA in CPT-seq until after dilution, which prevents the DNA from fragmenting. The

Tn5-bound DNA is then twice separated into pools and indexed such that each combinatorial index pool contains fragments with lengths summing to 5–10% of the total length of the genome. Thus, the regions sharing the same index are likely to be disjoint in the genome. All pools are combined and sequenced, resulting in reads with two barcodes. Two reads with the same barcodes are much more likely to have come from the same starting molecule than two reads without the same barcodes. The scaffolder fragScaff (146) can then use this information to scaffold contigs. CPT-seq does not create chromosome-scale scaffolds, but it is useful for assembling scaffolds big enough to further scaffold with other techniques, such as Hi-C.

Synteny: genes on the same chromosome

The 10x Genomics process uses a microfluidic system to create linked reads (**Figure 4*b***). First, this process creates small droplets, each consisting of 4–6 molecules of DNA with length in the tens or hundreds of kilobases, a gel bead containing millions of copies of a barcoded primer, and reagents necessary for the first steps of Illumina library preparation. Then, the gel beads are dissolved, releasing the barcoded primers into the droplets. The DNA is then amplified inside the droplets with the barcoded primers and recovered in solution, after which library preparation is completed. When the libraries are sequenced, each read contains a barcode identifying its source droplet. This linkage among sets of reads can then be used to phase haplotypes (147), identify structural variants (148), or assemble genomes (149). Supernova, 10x's assembler, can assemble diploid reference genomes, resolving structural variants between haplotypes (149). Advantages to linked-read sequencing include lower cost than long reads; a smaller input DNA requirement (∼1 ng); and its use of Illumina sequencers, which are already widespread (147).

Optical Maps

Several new sequencing-free high-throughput technologies use fluorescent labels to generate long-range information about a genome. One of these approaches, commercialized by Bionano, uses a nicking endonuclease to nick large fragments of DNA and then fluorescently labels the nicking sites (150). The fluorescently labeled DNA fragments are then electrophoretically fed through a nanochannel array and imaged to determine the sizes of the molecules and the locations of fluorescent labels. This information is assembled into a genome map, which can then be used to find structural variants (150, 151) or to scaffold contigs (152, 153), as shown in **Figure 4*c***. Another approach, commercialized by Opgen, generates restriction maps. Restriction digestion is performed on high–molecular weight DNA in situ on an optical mapping surface. Restriction fragment lengths are measured in situ for each DNA fragment using a DNA fluorescent dye. This results in an ordered restriction map giving the distances between restriction sites along the original molecule, to which the restriction sequences on contigs can be aligned. Opgen has been used in an assembly of the domestic goat genome (154).

Synteny-Based Methods

Another source of information that can be used to scaffold a genome assembly is comparison to the genome organization of a related species (**Figure 4*d***). Current software packages designed for assembly of bacterial genomes based on synteny include Ragout (155) and GAAP (156). Ragout relies on the insight, first implemented in the program RACA (157), that using information from multiple related species instead of just one can improve assembly accuracy by reducing the bias caused by rearrangements specific to a single reference genome. Ragout (155) improves on RACA by using phylogeny-guided multiple alignments of many reference genomes as a guide instead of a single reference genome with outgroups as additional information. The most recent version of Ragout (158) can also scaffold mammalian genomes. GAAP (156) uses a set of core bacterial genes

that are highly conserved and less likely to move around the genome as anchors to build scaffolds around.

The great advantage of synteny-based assembly methods is that they do not require collecting new data but instead rely on existing reference genomes. The primary disadvantages to these methods are that some lineages, such as mammals (159), have more structural rearrangement among species than others, such as archosaurs (50), and that synteny-based methods require the existence of at least one chromosome-scale assembly of a closely related organism, although the latter continuously becomes less of an issue as more chromosome-scale assemblies are published each year.

Trio Binning

Sequencing a trio of two parents and a child is a common method for haplotype phasing, because each chromosome in the offspring's genome came fully from either the mother or the father (160). A new method, trio binning, can assemble a diploid reference genome using short reads from the parental genomes and long reads from the offspring's genome. The short reads from the parents are aligned to the long reads from the offspring to divide the long reads into two sets: those from the maternal and those from the paternal haplotypes. Then, the two sets of long reads are assembled independently to create two separate haploid genomes (161). Although this method requires fully sequencing three individuals instead of just one, as in most other assembly techniques, it avoids the difficulties associated with assembling diploid genomes by breaking the problem into two haploid assemblies.

FUTURE CHALLENGES

Many new technologies can now be used to create chromosome-scale assemblies without costly and time-consuming methods such as BAC-end sequencing and physical mapping. Each of these new methods has its own strengths and weaknesses, so in practice, most chromosome-scale assembly projects today leverage the strengths of different data sources to construct the best assembly possible. Many recent projects have used various combinations of data types, such as long and short reads (87, 162, 163), long and/or short reads with proximity ligation (50, 89, 164), synteny and optical mapping (165), and short and linked reads (166, 167).

Perhaps the foremost challenge presented by the advent of these new technologies is determining how best to integrate them. Although there are clear precedents for how to best use individual data types, such as de Bruijn graph assembly for short reads, assembly projects using multiple data types must use ad hoc approaches to chain different pieces of software together into a pipeline. This problem is exacerbated by the fact that many software packages designed for specific data types are proprietary and/or closed source, making them harder to integrate into longer assembly pipelines.

Other challenges posed by these new technologies are more logistical and specific to individual data types. For example, Hi-C library preparation requires a large number of intact cells from the target organism, which can be harder to obtain, store, and transport than purified DNA for some organisms. Trio binning requires parent identification and sequencing, which is not feasible for some organisms, such as marine invertebrates that reproduce by releasing sperm or eggs into the water. Long reads are expensive to produce and have high error rates, although the technologies continue to improve. However, with the large number of choices available, there exist combinations of methods for chromosome-scale assembly and scaffolding that can meet the needs of most of today's genome projects.

DISCLOSURE STATEMENT

R.E.G. is a cofounder and paid consultant of Dovetail Genomics LLC.

ACKNOWLEDGMENTS

E.S.R. is supported by a Baskin School of Engineering Dissertation Year Fellowship and an ARCS Foundation Fellowship. R.E.G. is supported in part by National Institutes of Health grant U24HG009084. We thank Joel Armstrong for assistance obtaining the data used in **Figure 1** and Nathan K. Schaefer for critical reading of the manuscript.

LITERATURE CITED

1. Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, et al. 2000. The genome sequence of *Drosophila melanogaster*. *Science* 287:2185–95
2. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, et al. 2001. The sequence of the human genome. *Science* 291:1304–51
3. Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, et al. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* 420:520
4. Sanger F, Nicklen S, Coulson AR. 1977. DNA sequencing with chain-terminating inhibitors. *PNAS* 74:5463–67
5. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, et al. 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437:376–80
6. McKernan KJ, Peckham HE, Costa GL, McLaughlin SF, Fu Y, et al. 2009. Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. *Genome Res.* 19:1527–41
7. Rothberg JM, Hinz W, Rearick TM, Schultz J, Mileski W, et al. 2011. An integrated semiconductor device enabling non-optical genome sequencing. *Nature* 475:348–52
8. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, et al. 2008. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* 456:53–59
9. Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, et al. 2008. The complete genome of an individual by massively parallel DNA sequencing. *Nature* 452:872–76
10. Metzker ML. 2010. Sequencing technologies—the next generation. *Nat. Rev. Genet.* 11:31
11. Stein LD. 2010. The case for cloud computing in genome informatics. *Genome Biol.* 11:207
12. Glenn TC. 2011. Field guide to next-generation DNA sequencers. *Mol. Ecol. Resour.* 11:759–69
13. Heather JM, Chain B. 2016. The sequence of sequencers: the history of sequencing DNA. *Genomics* 107:1–8
14. Suh A, Weber CC, Kehlmaier C, Braun EL, Green RE, et al. 2014. Early Mesozoic coexistence of amniotes and Hepadnaviridae. *PLOS Genet.* 10:e1004559
15. Cahill JA, Heintzman PD, Harris K, Teasdale MD, Kapp J, et al. 2018. Genomic evidence of widespread admixture from polar bears into brown bears during the last ice age. *Mol. Biol. Evol.* 35:1120–29
16. Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, et al. 2010. A draft sequence of the Neandertal genome. *Science* 328:710–22
17. Prüfer K, Racimo F, Patterson N, Jay F, Sankararaman S, et al. 2014. The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature* 505:43–49
18. Prüfer K, de Filippo C, Grote S, Mafessoni F, Korlević P, et al. 2017. A high-coverage Neandertal genome from Vindija Cave in Croatia. *Science* 358:655–58
19. Lamichhaney S, Berglund J, Almén MS, Maqbool K, Grabherr M, et al. 2015. Evolution of Darwin's finches and their beaks revealed by genome sequencing. *Nature* 518:371–75
20. Brandvain Y, Kenney AM, Flagel L, Coop G, Sweigart AL. 2014. Speciation and introgression between *Mimulus nasutus* and *Mimulus guttatus*. *PLOS Genet.* 10:e1004410
21. Hsu PD, Lander ES, Zhang F. 2014. Development and applications of CRISPR-Cas9 for genome engineering. *Cell* 157:1262–78

22. Jinek M, Chylinski K, Fonfara I, Hauer M, Doudna JA, Charpentier E. 2012. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* 337:816–21
23. Shah AN, Davey CF, Whitebirch AC, Miller AC, Moens CB. 2015. Rapid reverse genetic screening using CRISPR in zebrafish. *Nat. Methods* 12:535–40
24. Gurumurthy CB, Grati M, Ohtsuka M, Schilit SLP, Quadros RM, Liu XZ. 2016. CRISPR: a versatile tool for both forward and reverse genetics research. *Hum. Genet.* 135:971–76
25. Schatz MC, Delcher AL, Salzberg SL. 2010. Assembly of large genomes using second-generation sequencing. *Genome Res.* 20:1165–73
26. Ye L, Hillier LW, Minx P, Thane N, Locke DP, et al. 2011. A vertebrate case study of the quality of assemblies derived from next-generation sequences. *Genome Biol.* 12:R31
27. Alkan C, Sajjadian S, Eichler EE. 2011. Limitations of next-generation genome sequence assembly. *Nat. Methods* 8:61–65
28. Lewin HA, Larkin DM, Pontius J, O'Brien SJ. 2009. Every genome sequence needs a good map. *Genome Res.* 19:1925–28
29. Minoche AE, Dohm JC, Himmelbauer H. 2011. Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and genome analyzer systems. *Genome Biol.* 12:R112
30. Eid J, Fehr A, Gray J, Luong K, Lyle J, et al. 2009. Real-time DNA sequencing from single polymerase molecules. *Science* 323:133–38
31. Rhoads A, Au KF. 2015. PacBio sequencing and its applications. *Genom. Proteom. Bioinform.* 13:278–89
32. Jain M, Olsen HE, Paten B, Akeson M. 2016. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biol.* 17:239
33. Meyer M, Kircher M. 2010. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harb. Protoc.* 2010(6):pdb.prot5448
34. Picelli S, Björklund AK, Reinius B, Sagasser S, Winberg G, Sandberg R. 2014. Tn5 transposase and tagmentation procedures for massively scaled sequencing projects. *Genome Res.* 24:2033–40
35. Gansauge M-T, Gerber T, Glocke I, Korlevic P, Lippik L, et al. 2017. Single-stranded DNA library preparation from highly degraded DNA using T4 DNA ligase. *Nucleic Acids Res.* 45:e79
36. Lu H, Giordano F, Ning Z. 2016. Oxford Nanopore MinION sequencing and genome assembly. *Genom. Proteom. Bioinform.* 14:265–79
37. Mayjonade B, Gouzy J, Donnadieu C, Pouilly N, Marande W, et al. 2016. Extraction of high-molecular-weight genomic DNA for long-read sequencing of single molecules. *Biotechniques* 61:203–5
38. Staden R. 1979. A strategy of DNA sequencing employing computer programs. *Nucleic Acids Res.* 6:2601–10
39. Li Z, Chen Y, Mu D, Yuan J, Shi Y, et al. 2012. Comparison of the two major classes of assembly algorithms: overlap-layout-consensus and de-Bruijn-graph. *Brief. Funct. Genom.* 11:25–37
40. Compeau PEC, Pevzner PA, Tesler G. 2011. How to apply de Bruijn graphs to genome assembly. *Nat. Biotechnol.* 29:987–91
41. Simpson JT, Pop M. 2015. The theory and practice of genome sequence assembly. *Annu. Rev. Genom. Hum. Genet.* 16:153–72
42. Britten RJ, Kohne DE. 1968. Repeated sequences in DNA. *Science* 161:529–40
43. Treangen TJ, Salzberg SL. 2012. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat. Rev. Genet.* 13:36–46
44. Kim J, Farre M, Auvil L, Capitanu B, Larkin DM, et al. 2017. Reconstruction and evolutionary history of eutherian chromosomes. *PNAS* 114:E5379–88
45. Larkin DM, Pape G, Donthu R, Auvil L, Welge M, Lewin HA. 2009. Breakpoint regions and homologous synteny blocks in chromosomes have different evolutionary histories. *Genome Res.* 19:770–77
46. Sagai T, Hosoya M, Mizushina Y, Tamura M, Shiroishi T. 2005. Elimination of a long-range *cis*-regulatory module causes complete loss of limb-specific *Shh* expression and truncation of the mouse limb. *Development* 132:797–803
47. Maston GA, Evans SK, Green MR. 2006. Transcriptional regulatory elements in the human genome. *Annu. Rev. Genom. Hum. Genet.* 7:29–59

48. Pombo A, Dillon N. 2015. Three-dimensional genome architecture: players and mechanisms. *Nat. Rev. Mol. Cell Biol.* 16:245–57
49. Batugedara G, Lu XM, Bunnik EM, Le Roch KG. 2017. The role of chromatin structure in gene regulation of the human malaria parasite. *Trends Parasitol.* 33:364–77
50. Rice ES, Kohno S, John JS, Pham S, Howard J, et al. 2017. Improved genome assembly of American alligator genome reveals conserved architecture of estrogen signaling. *Genome Res.* 27:686–96
51. Jensen-Seaman MI, Furey TS, Payseur BA, Lu Y, Roskin KM, et al. 2004. Comparative recombination rates in the rat, mouse, and human genomes. *Genome Res.* 14:528–38
52. Backström N, Forstmeier W, Schielzeth H, Mellenius H, Nam K, et al. 2010. The recombination landscape of the zebra finch *Taeniopygia guttata* genome. *Genome Res.* 20:485–95
53. Lichten M, Goldman ASH. 1995. Meiotic recombination hotspots. *Annu. Rev. Genet.* 29:423–44
54. Ellegren H. 2010. Evolutionary stasis: the stable chromosomes of birds. *Trends Ecol. Evol.* 25:283–91
55. Murray GGR, Soares AER, Novak BJ, Schaefer NK, Cahill JA, et al. 2017. Natural selection shaped the rise and fall of passenger pigeon genomic diversity. *Science* 358:951–54
56. Lewis CM, Knight J. 2012. Introduction to genetic association studies. *Cold Spring Harb. Protoc.* 2012(3):297–306
57. Lohmueller KE, Pearce CL, Pike M, Lander ES, Hirschhorn JN. 2003. Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. *Nat. Genet.* 33:177
58. Iles MM. 2008. What can genome-wide association studies tell us about the genetics of common disease? *PLOS Genet.* 4:e33
59. Chang D, Nalls MA, Hallgrímsdóttir IB, Hunkapiller J, van der Brug M, et al. 2017. A meta-analysis of genome-wide association studies identifies 17 new Parkinson's disease risk loci. *Nat. Genet.* 49:1511–16
60. Lee JC, Biasci D, Roberts R, Gearry RB, Mansfield JC, et al. 2017. Genome-wide association study identifies distinct genetic contributions to prognosis and susceptibility in Crohn's disease. *Nat. Genet.* 49:262–68
61. Parchman TL, Gompert Z, Mudge J, Schilkey FD, Benkman CW, Buerkle CA. 2012. Genome-wide association genetics of an adaptive trait in lodgepole pine. *Mol. Ecol.* 21:2991–3005
62. Cai Q, Qian X, Lang Y, Luo Y, Xu J, et al. 2013. Genome sequence of ground tit *Pseudopodoces humilis* and its adaptation to high altitude. *Genome Biol.* 14:R29
63. Cohen MM, Gans C. 1970. The chromosomes of the order Crocodilia. *Cytogenetics* 9:81–105
64. Srikulnath K, Thapana W, Muangmai N. 2015. Role of chromosome changes in *Crocodylus* evolution and diversity. *Genom. Inform.* 13:102–11
65. Murphy WJ, Larkin DM, Everts-van der Wind A, Bourque G, Tesler G, et al. 2005. Dynamics of mammalian chromosome evolution inferred from multispecies comparative maps. *Science* 309:613–17
66. Graphodatsky AS, Trifonov VA, Stanyon R. 2011. The genome diversity and karyotype evolution of mammals. *Mol. Cytogenet.* 4:22
67. Perelman P, Johnson WE, Roos C, Seuánez HN, Horvath JE, et al. 2011. A molecular phylogeny of living primates. *PLOS Genet.* 7:e1001342
68. Noor MAF, Grams KL, Bertucci LA, Reiland J. 2001. Chromosomal inversions and the reproductive isolation of species. *PNAS* 98:12084–88
69. Walsh JB. 1982. Rate of accumulation of reproductive isolation by chromosome rearrangements. *Am. Nat.* 120:510–32
70. Abbott JK, Nordén AK, Hansson B. 2017. Sex chromosome evolution: historical insights and future perspectives. *Proc. R. Soc. B* 284:20162806
71. Salse J, Abrouk M, Bolot S, Guilhot N, Courcelle E, et al. 2009. Reconstruction of monocotelydoneous proto-chromosomes reveals faster evolution in plants than in animals. *PNAS* 106:14908–13
72. Murat F, Zhang R, Guizard S, Gavranović H, Flores R, et al. 2015. Karyotype and gene order evolution from reconstructed extinct ancestors highlight contrasts in genome plasticity of modern rosid crops. *Genome Biol. Evol.* 7:735–49

73. O'Connor RE, Romanov MN, Kiazim LG, Barrett PM, Farré M, et al. 2018. Reconstruction of the diapsid ancestral genome permits chromosome evolution tracing in avian and non-avian dinosaurs. *Nat. Commun.* 9:1883
74. Morgan TH. 1911. Random segregation versus coupling in Mendelian inheritance. *Science* 34:384
75. Sturtevant AH. 1913. The linear arrangement of six sex-linked factors in *Drosophila*, as shown by their mode of association. *J. Exp. Zool.* 14:43–59
76. Mascher M, Stein N. 2014. Genetic anchoring of whole-genome shotgun assemblies. *Front. Genet.* 5:208
77. Fierst JL. 2015. Using linkage maps to correct and scaffold *de novo* genome assemblies: methods, challenges, and computational tools. *Front. Genet.* 6:220
78. Mascher M, Stein N. 2014. Genetic anchoring of whole-genome shotgun assemblies. *Front. Genet.* 5:208
79. Nossa CW, Havlak P, Yue J-X, Lv J, Vincent KY, et al. 2014. Joint assembly and genetic mapping of the Atlantic horseshoe crab genome reveals ancient whole genome duplication. *Gigascience* 3:9
80. Kai W, Kikuchi K, Tohari S, Chew AK, Tay A, et al. 2011. Integration of the genetic map and genome assembly of fugu facilitates insights into distinct features of genome evolution in teleosts and mammals. *Genome Biol. Evol.* 3:424–42
81. Goss SJ, Harris H. 1975. New method for mapping genes in human chromosomes. *Nature* 255:680
82. Cox DR, Burmeister M, Price ER, Kim S, Myers RM. 1990. Radiation hybrid mapping: a somatic cell genetic method for constructing high-resolution maps of mammalian chromosomes. *Science* 250:245–50
83. Deloukas P. 2005. Radiation hybrid mapping. *eLS*. **https://doi.org/10.1038/npg.els.0005361**
84. Kumar A, Bassi FM, Paux E, Al-Azzam O, de Jimenez MM, et al. 2012. DNA repair and crossing over favor similar chromosome regions as discovered in radiation hybrid of *Triticum*. *BMC Genom.* 13:339
85. Yang Y-P, Womack JE. 1998. Parallel radiation hybrid mapping: a powerful tool for high-resolution genomic comparison. *Genome Res.* 8:731–36
86. Howe K, Clark MD, Torroja CF, Torrance J, Berthelot C, et al. 2013. The zebrafish reference genome sequence and its relationship to the human genome. *Nature* 496:498
87. Bickhart DM, Rosen BD, Koren S, Sayre BL, Hastie AR, et al. 2017. Single-molecule sequencing and chromatin conformation capture enable *de novo* reference assembly of the domestic goat genome. *Nat. Genet.* 49:643
88. Warren WC, Hillier LW, Tomlinson C, Minx P, Kremitzki M, et al. 2017. A new chicken genome assembly provides insight into avian genome structure. *G3 Genes Genomes Genet.* 7:109–17
89. Kalbfleisch TS, Rice E, DePriest MS, Walenz BP, Hestand MS, et al. 2018. EquCab3, an updated reference genome for the domestic horse. bioRxiv 306928. **https://doi.org/10.1101/306928**
90. Espinosa R, Le Beau MM. 2000. Gene mapping by FISH. In *The Nucleic Acid Protocols Handbook*, ed. R Rapley, pp. 991–1010. New York: Springer
91. Speicher MR, Ballard SG, Ward DC. 1996. Karyotyping human chromosomes by combinatorial multifluor FISH. *Nat. Genet.* 12:368
92. Schröck E, Du Manoir S, Veldman T, Schoell B, Wienberg J, et al. 1996. Multicolor spectral karyotyping of human chromosomes. *Science* 273:494–97
93. Fan Y-S, Davis LM, Shows TB. 1990. Mapping small DNA sequences by fluorescence in situ hybridization directly on banded metaphase chromosomes. *PNAS* 87:6223–27
94. Trask B, Pinkel D, van den Engh G. 1989. The proximity of DNA sequences in interphase cell nuclei is correlated to genomic distance and permits ordering of cosmids spanning 250 kilobase pairs. *Genomics* 5:710–17
95. Raap AK, Florijn RJ, Blonden LAJ, Wiegant J, Vaandrager J-W, et al. 1996. Fiber FISH as a DNA mapping tool. *Methods* 9:67–73
96. Shearer LA, Anderson LK, De Jong H, Smit S, Goicoechea JL, et al. 2014. Fluorescence in situ hybridization and optical mapping to correct scaffold arrangement in the tomato genome. *G3 Genes Genomes Genet.* 4:1395–405
97. Vij S, Kuhl H, Kuznetsova IS, Komissarov A, Yurchenko AA, et al. 2016. Chromosomal-level assembly of the Asian seabass genome using long sequence reads and multi-layered scaffolding. *PLOS Genet.* 12:e1005954

98. O'Connor M, Peifer M, Bender W. 1989. Construction of large DNA segments in *Escherichia coli*. *Science* 244:1307–12
99. Kelley JM, Field CE, Craven MB, Rounsley SD, Adams MD, et al. 1999. High throughput direct end sequencing of BAC clones. *Nucleic Acids Res*. 27:1539–46
100. Han CS, Sutherland RD, Jewett PB, Campbell ML, Meincke LJ, et al. 2000. Construction of a BAC contig map of chromosome 16q by two-dimensional overgo hybridization. *Genome Res*. 10:714–21
101. Ardui S, Ameur A, Vermeesch JR, Hestand MS. 2018. Single molecule real-time (SMRT) sequencing comes of age: applications and utilities for medical diagnostics. *Nucleic Acids Res* 46:2159–68
102. Levene MJ, Korlach J, Turner SW, Foquet M, Craighead HG, Webb WW. 2003. Zero-mode waveguides for single-molecule analysis at high concentrations. *Science* 299:682–86
103. Chin C-S, Peluso P, Sedlazeck FJ, Nattestad M, Concepcion GT, et al. 2016. Phased diploid genome assembly with single-molecule real-time sequencing. *Nat. Methods* 13:1050–54
104. Ritz A, Bashir A, Sindi S, Hsu D, Hajirasouliha I, Raphael BJ. 2014. Characterization of structural variants with single molecule and hybrid sequencing approaches. *Bioinformatics* 30:3458–66
105. Feng Y, Zhang Y, Ying C, Wang D, Du C. 2015. Nanopore-based fourth-generation DNA sequencing technology. *Genom. Proteom. Bioinform.* 13:4–16
106. Cherf GM, Lieberman KR, Rashid H, Lam CE, Karplus K, Akeson M. 2012. Automated forward and reverse ratcheting of DNA in a nanopore at 5-Å precision. *Nat. Biotechnol.* 30:344–48
107. Jain M, Koren S, Miga KH, Quick J, Rand AC, et al. 2018. Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat. Biotechnol.* 36:338–45
108. Payne A, Holmes N, Rakyan V, Loose M. 2018. Whale watching with BulkVis: a graphical viewer for Oxford Nanopore bulk fast5 files. bioRxiv 312256. **https://doi.org/10.1101/312256**
109. Laver T, Harrison J, O'Neill PA, Moore K, Farbos A, et al. 2015. Assessing the performance of the Oxford Nanopore Technologies MinION. *Biomol. Detect. Quantif.* 3:1–8
110. Wick RR, Judd LM, Holt KE. 2018. Comparison of Oxford Nanopore basecalling tools. **https://doi.org/10.5281/zenodo.1188469**
111. Madoui M-A, Engelen S, Cruaud C, Belser C, Bertrand L, et al. 2015. Genome assembly using Nanopore-guided long and error-free DNA reads. *BMC Genom*. 16:327
112. Sović I, Križanović K, Skala K, Šikić M. 2016. Evaluation of hybrid and non-hybrid methods for *de novo* assembly of nanopore reads. *Bioinformatics* 32:2582–89
113. Tyson JR, O'Neil NJ, Jain M, Olsen HE, Hieter P, Snutch TP. 2018. MinION-based long-read sequencing and assembly extends the *Caenorhabditis elegans* reference genome. *Genome Res*. 28:266–74
114. Giordano F, Aigrain L, Quail MA, Coupland P, Bonfield JK, et al. 2017. *De novo* yeast genome assemblies from MinION, PacBio and MiSeq platforms. *Sci. Rep.* 7:3935
115. Loman NJ, Quick J, Simpson JT. 2015. A complete bacterial genome assembled *de novo* using only nanopore sequencing data. *Nat. Methods* 12:733–35
116. Li C, Chng KR, Boey EJH, Ng AHQ, Wilm A, Nagarajan N. 2016. INC-Seq: accurate single molecule reads using nanopore sequencing. *Gigascience* 5:34
117. Ekblom R, Wolf JBW. 2014. A field guide to whole-genome sequencing, assembly and annotation. *Evol. Appl.* 7:1026–42
118. Koren S, Schatz MC, Walenz BP, Martin J, Howard JT, et al. 2012. Hybrid error correction and *de novo* assembly of single-molecule sequencing reads. *Nat. Biotechnol.* 30:693–700
119. Myers EW, Sutton GG, Delcher AL, Dew IM, Fasulo DP, et al. 2000. A whole-genome assembly of *Drosophila*. *Science* 287:2196–204
120. Lee H, Gurtowski J, Yoo S, Marcus S, McCombie WR, Schatz M. 2014. Error correction and assembly complexity of single molecule sequencing reads. bioRxiv 006395. **https://doi.org/10.1101.006395**
121. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, et al. 2004. Versatile and open software for comparing large genomes. *Genome Biol.* 5:R12
122. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, et al. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* 19:455–77

123. Ye C, Hill CM, Wu S, Ruan J, Ma Z. 2016. DBG2OLC: efficient assembly of large genomes using long erroneous reads of the third generation sequencing technologies. *Sci. Rep.* 6:31900
124. Chin C-S, Alexander DH, Marks P, Klammer AA, Drake J, et al. 2013. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat. Methods* 10:563
125. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. 2017. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res.* 27:722–36
126. Berlin K, Koren S, Chin C-S, Drake JP, Landolin JM, Phillippy AM. 2015. Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. *Nat. Biotechnol.* 33:623–30
127. Miller JR, Delcher AL, Koren S, Venter E, Walenz BP, et al. 2008. Aggressive assembly of pyrosequencing reads with mates. *Bioinformatics* 24:2818–24
128. Li H. 2016. Minimap and miniasm: fast mapping and *de novo* assembly for noisy long sequences. *Bioinformatics* 32:2103–10
129. Kolmogorov M, Yuan J, Lin Y, Pevzner P. 2018. Assembly of long error-prone reads using repeat graphs. bioRxiv 247148. **https://doi.org/10.1101/247148**
130. Lin Y, Yuan J, Kolmogorov M, Shen MW, Chaisson M, Pevzner PA. 2016. Assembly of long error-prone reads using de Bruijn graphs. *PNAS* 113:E8396–E405
131. English AC, Richards S, Han Y, Wang M, Vee V, et al. 2012. Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology. *PLOS ONE* 7:e47768
132. Warren RL, Yang C, Vandervalk BP, Behsaz B, Lagman A, et al. 2015. LINKS: scalable, alignment-free scaffolding of draft genomes with long reads. *Gigascience* 4:35
133. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, et al. 2014. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLOS ONE* 9:e112963
134. de Wit E, de Laat W. 2012. A decade of 3C technologies: insights into nuclear organization. *Genes Dev.* 26:11–24
135. Dekker J, Rippe K, Dekker M, Kleckner N. 2002. Capturing chromosome conformation. *Science* 295:1306–11
136. Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, et al. 2009. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* 326:289–93
137. Burton JN, Adey A, Patwardhan RP, Qiu R, Kitzman JO, Shendure J. 2013. Chromosome-scale scaffolding of *de novo* genome assemblies based on chromatin interactions. *Nat. Biotechnol.* 31:1119–25
138. Putnam NH, O'Connell BL, Stites JC, Rice BJ, Blanchette M, et al. 2016. Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. *Genome Res.* 26:342–50
139. Kaplan N, Dekker J. 2013. High-throughput genome scaffolding from *in vivo* DNA interaction frequency. *Nat. Biotechnol.* 31:1143
140. Korbel JO, Lee C. 2013. Genome assembly and haplotyping with Hi-C. *Nat. Biotechnol.* 31:1099
141. Dudchenko O, Batra SS, Omer AD, Nyquist SK, Hoeger M, et al. 2017. *De novo* assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science* 356:92–95
142. Dudchenko O, Shamim MS, Batra S, Durand NC, Musial NT, et al. 2018. The Juicebox Assembly Tools module facilitates *de novo* assembly of mammalian genomes with chromosome-length scaffolds for under $1000. bioRxiv 254797. **https://doi.org/10.1101/254797**
143. Ghurye J, Pop M, Koren S, Bickhart D, Chin C-S. 2017. Scaffolding of long read assemblies using long range contact information. *BMC Genom.* 18:527
144. Ghurye J, Rhie A, Walenz BP, Schmitt A, Selvaraj S, et al. 2018. Integrating Hi-C links with assembly graphs for chromosome-scale assembly. bioRxiv 261149. **https://doi.org/10.1101/261149**
145. Amini S, Pushkarev D, Christiansen L, Kostem E, Royce T, et al. 2014. Haplotype-resolved whole-genome sequencing by contiguity-preserving transposition and combinatorial indexing. *Nat. Genet.* 46:1343–49
146. Adey A, Kitzman JO, Burton JN, Daza R, Kumar A, et al. 2014. *In vitro*, long-range sequence information for *de novo* genome assembly via transposase contiguity. *Genome Res.* 24:2041–49
147. Zheng GXY, Lau BT, Schnall-Levin M, Jarosz M, Bell JM, et al. 2016. Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. *Nat. Biotechnol.* 34:303–11

148. Garcia S, Williams S, Xu AW, Herschleb J, Marks P, et al. 2017. Linked-read sequencing resolves complex structural variants. bioRxiv 231662. **https://doi.org/10.1101/231662**
149. Weisenfeld NI, Kumar V, Shah P, Church DM, Jaffe DB. 2017. Direct determination of diploid genome sequences. *Genome Res.* 27:757–67
150. Mak ACY, Lai YYY, Lam ET, Kwok T-P, Leung AKY, et al. 2016. Genome-wide structural variation detection by genome mapping on nanochannel arrays. *Genetics* 202:351–62
151. Cao H, Hastie AR, Cao D, Lam ET, Sun Y, et al. 2014. Rapid detection of structural variation in a human genome using nanochannel-based genome mapping technology. *Gigascience* 3:34
152. Shi L, Guo Y, Dong C, Huddleston J, Yang H, et al. 2016. Long-read sequencing and *de novo* assembly of a Chinese genome. *Nat. Commun.* 7:12065
153. Seo J-S, Rhie A, Kim J, Lee S, Sohn M-H, et al. 2016. *De novo* assembly and phasing of a Korean human genome. *Nature* 538:243–47
154. Dong Y, Xie M, Jiang Y, Xiao N, Du X, et al. 2013. Sequencing and automated whole-genome optical mapping of the genome of a domestic goat (*Capra hircus*). *Nat. Biotechnol.* 31:135–41
155. Kolmogorov M, Raney B, Paten B, Pham S. 2014. Ragout-a reference-assisted assembly tool for bacterial genomes. *Bioinformatics* 30:i302–9
156. Yuan L, Yu Y, Zhu Y, Li Y, Li C, et al. 2017. GAAP: Genome-organization-framework-Assisted Assembly Pipeline for prokaryotic genomes. *BMC Genom.* 18:952
157. Kim J, Larkin DM, Cai Q, Zhang Y, Ge R-L, et al. 2013. Reference-assisted chromosome assembly. *PNAS* 110:1785–90
158. Kolmogorov M, Armstrong J, Raney BJ, Streeter I, Dunn M, et al. 2018. Chromosome assembly of large and complex genomes using multiple references. bioRxiv 088435. **https://doi.org/10.1101/088435**
159. Murphy WJ, Larkin DM, Everts-van der Wind A, Bourque G, Tesler G, et al. 2005. Dynamics of mammalian chromosome evolution inferred from multispecies comparative maps. *Science* 309:613–17
160. Glusman G, Cox HC, Roach JC. 2014. Whole-genome haplotyping approaches and genomic medicine. *Genome Med.* 6:73
161. Koren S, Rhie A, Walenz BP, Dilthey AT, Bickhart DM, et al. 2018. Complete assembly of parental haplotypes with trio binning. bioRxiv 271486. **https://doi.org/10.1101.271486**
162. Gordon D, Huddleston J, Chaisson MJP, Hill CM, Kronenberg ZN, et al. 2016. Long-read sequence assembly of the gorilla genome. *Science* 352:aae0344
163. Larsen PA, Harris RA, Liu Y, Murali SC, Campbell CR, et al. 2017. Hybrid *de novo* genome assembly and centromere characterization of the gray mouse lemur (*Microcebus murinus*). *BMC Biol.* 15:110
164. Kuderna LFK, Tomlinson C, Hillier LW, Tran A, Fiddes IT, et al. 2017. A 3-way hybrid approach to generate a new high-quality chimpanzee reference genome (Pan_tro_3.0). *Gigascience* 6:1–6
165. Thybert D, Roller M, Navarro FCP, Fiddes I, Streeter I, et al. 2018. Repeat associated mechanisms of genome evolution and function revealed by the *Mus caroli* and *Mus pahari* genomes. *Genome Res.* 28:448–59
166. Hammond SA, Warren RL, Vandervalk BP, Kucuk E, Khan H, et al. 2017. The North American bullfrog draft genome provides insight into hormonal regulation of long noncoding RNA. *Nat. Commun.* 8:1433
167. Jones SJM, Taylor GA, Chan S, Warren RL, Hammond SA, et al. 2017. The genome of the beluga whale (*Delphinapterus leucas*). *Genes* 8:378
168. Zerbino DR, Birney E. 2008. Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome Res.* 18:821–29
169. Luo R, Liu B, Xie Y, Li Z, Huang W, et al. 2012. SOAPdenovo2: an empirically improved memory-efficient short-read *de novo* assembler. *Gigascience* 1:18
170. Chapman JA, Ho I, Sunkara S, Luo S, Schroth GP, Rokhsar DS. 2011. Meraculous: *de novo* genome assembly with short paired-end reads. *PLOS ONE* 6:e23501
171. Gnerre S, MacCallum I, Przybylski D, Ribeiro FJ, Burton JN, et al. 2011. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *PNAS* 108:1513–18
172. Simpson JT, Durbin R. 2012. Efficient *de novo* assembly of large genomes using compressed data structures. *Genome Res.* 22:549–56

173. Jackman SD, Vandervalk BP, Mohamadi H, Chu J, Yeo S, et al. 2017. ABySS 2.0: resource-efficient assembly of large genomes using a Bloom filter. *Genome Res*. 27:768–77
174. Love RR, Weisenfeld NI, Jaffe DB, Besansky NJ, Neafsey DE. 2016. Evaluation of DISCOVAR *de novo* using a mosquito sample for cost-effective short-read genome assembly. *BMC Genom*. 17:187
175. Jarvis ED, Mirarab S, Aberer AJ, Li B, Houde P, et al. 2014. Whole-genome analyses resolve early branches in the tree of life of modern birds. *Science* 346:1320–31
176. Gordon D, Huddleston J, Chaisson MJ, Hill CM, Kronenberg ZN, et al. 2016. Long-read sequence assembly of the gorilla genome. *Science* 352:aae0344
177. Jain M, Koren S, Miga KH, Quick J, Rand AC, et al. 2018. Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat. Biotechnol.* 36:338–45