INVITED PAPER

Building and Improving Reference Genome Assemblies

This paper reviews the problems and algorithms of assembling a complete genome from millions of short DNA sequencing reads.

By Karyn Meltz Steinberg, Valerie A. Schneider, Can Alkan, Michael J. Montague, Wesley C. Warren, Deanna M. Church, and Richard K. Wilson

ABSTRACT | A genome sequence assembly provides the foundation for studies of genotypic and phenotypic variation, genome structure, and evolution of the target organism. In the past four decades, there has been a surge of new sequencing technologies, and with these developments, computational scientists have developed new algorithms to improve genome assembly. Here we discuss the relationship between sequencing technology improvements and assembly algorithm development and how these are applied to extend and improve human and nonhuman genome assemblies.

KEYWORDS | Bioinformatics; DNA; genetics; genomics

I. INTRODUCTION

The current genomics revolution has been driven by high throughput, low-cost sequencing. It is currently not easy, cost effective, or feasible to perform *de novo* assembly of large numbers of big, complex genomes. *De novo* assembly is analogous to putting together a jigsaw puzzle where the image is a blue sky with few clouds. Many of the pieces look exactly the same, and the box lid is missing so there is no template. The general approach for genome analysis has been to expend resources to develop a reference genome assembly which is used to support population level genome analysis. An organism's genome is a physical object, and the reference genome assembly is a representation or a model of that object. Reference-based analysis strategies allow scientists to align sequencing reads (use algorithms to take the output from DNA sequencing and compare to the assembly) and call variants (differences in the DNA sequence output from the assembly) relative to the reference. These "resequencing" strategies allow users to analyze more genomes at lower costs. However, this approach is limited by many factors. Typically, only limited regions of the genome are reliably accessible to resequencing methods. Additionally, reference genome assemblies are constantly evolving and improving, which is important as resequencing-based analysis is dependent upon having a high-quality reference assembly; however, most data sets are not reanalyzed on updated references, due to time and cost. In this review, we discuss the close relationship between sequencing technology and assembly algorithms and how these are used to improve reference assemblies.

DNA sequencing technologies have advanced significantly over the past four decades; however, we are still only able to sequence small segments of a genome at a time. Sequence reads are sequences generated by a machine from a DNA fragment, and read lengths range from short [150–250 base pairs (bp); e.g., sequencing by synthesis from Illumina) to medium (800–900 bp; e.g., dye terminator sequencing from Sanger) to large (>15 kilobase pairs (kbp) average; e.g., single molecule sequencing from Pacific Biosciences, PacBio] which is sufficient for most resequencing studies. Table 1 presents a summary of sequencing technologies discussed in this review. For *de novo* assembly, limited read lengths means that large complex eukaryotic genomes (such as the approximately 3 gigabase pair human genome) must be reconstructed from fragmented DNA sequences. The advantage of this method, known as shotgun sequencing, is that the genome can be fragmented and sequenced in parallel. Each fragment is random and is probabilistically expected to overlap another fragment such that, in theory, the entire genome can be assembled by comparing the similarity of the overlapping

Manuscript received December 22, 2015; revised October 17, 2016; accepted December 17, 2016. Date of publication January 27, 2017; date of current version February 16, 2017.

K. Meltz Steinberg and **W. C. Warren** are with the McDonnell Genome Institute, Washington University School of Medicine, St. Louis, MO, USA (e-mail: kmeltzst@genome.wustl.edu).
R. K. Wilson is with the Institute for Genomic Medicine, Nationwide Children's Hospital, Columbus, OH 43205 USA.
V. A. Schneider is with the National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD, USA.
C. Alkan is with the Department of Computer Engineering, Bilkent University, Bilkent, Ankara, Turkey.
M. J. Montague is with the Department of Neuroscience, University of Pennsylvania, Philadelphia, PA, USA.
D. M. Church is with Personalis, Inc. Menlo Park, CA, USA.

Digital Object Identifier: 10.1109/JPROC.2016.2645402

Table 1 Sequencing Technologies

Sequencing technology	Average Read length	Major error mode	Output per run
Sanger	700-900 bp	Polymerase slippage at homopolymers and short repeats	900 kb
Illumina	150-250 bp	Single base substitutions	150-300 Gb for 2500 1.8 Tb for HiSeq X Ten
Pacific Biosciences (PacBio)	10-15 kb	Indel substitutions	500 Mb-1 Gb
Oxford Nanopore (ONT) MinION	6 kb	Deletions	~500 Mb

sequences and pasting these together into increasingly larger, contiguous sequences called contigs.

The relationship between read length and the accuracy of assembly algorithms was statistically derived by Lander and Waterman in 1988 [1]. These equations estimate the size and number of genomic sequences that can be assembled from sequence reads based on read length, genome coverage, and the overlap between two reads that can be accurately detected by the assembly algorithm. Longer reads with more overlapping sequences are easier to assemble into larger contigs, while shorter reads are more difficult to assemble because there are often gaps between reads. For example, Alkan *et al.* [2] assessed multiple human genome assemblies and determined that *de novo* assemblies generated from short reads were significantly shorter than the human reference genome assembly (generated using a clone-based assembly approach from reads sequencing with Sanger-based technology) because of missing or collapsed repeat elements, sequences that are similar or identical to other sequences elsewhere in the genome. In a clone-based assembly, a set of large insert clones such as bacterial artificial chromosome (BAC) clones (in which fragments of DNA are inserted into the bacterial vector that is approximately 100–200 kbp) are individually sequenced and assembled. On the other hand, short-read assemblies are created from sequencing a library of DNA fragments that are often smaller than many repetitive genomic elements and, therefore, these assemblies are often more fragmented.

Hence, the major challenge for *de novo* assembly results from repetitive sequences. For example, almost 50% of the human genome is characterized by repetitive elements such as tandem and interspersed repeats, and segmental duplications. Tandem repeats are sequence elements with two to many thousands of copies that are physically adjacent, such as telomeres and centromeres. Interspersed repeats can be short (100–300 bp, e.g., *Alu* repeats) or long (>300 bp, e.g., L1 repeats) repeats that are not physically next to each other. Segmental duplications can be either tandem or interspersed and are defined as repeats greater than 1 kbp with greater than 90% homology to one another [3]. Some organisms, such as plants, can have more than 80% repetitive content in their genomes from elements such as transposons, and some have even undergone whole genome duplications [4], [5].

The presence of repetitive elements can result in gaps and/or collapse of sequence information in the assembly (Fig. 1), due to the fact that the algorithm cannot distinguish two identical or near-identical sequences. An assembly algorithm might resolve a repetitive element if its length is shorter than the read length; however, most repetitive elements do not meet this criterion when using standard short-read sequence lengths, and the cost of producing long-read sequence lengths is too high for routine, large-scale projects. More often, the repeat-flanking sequences become misassembled or incorrectly gapped, and if the repeat is tandem, the assembler will likely collapse the actual number of repetitive elements into fewer copies.

Genome assemblies contain gaps where algorithms cannot assemble sequence. There are three types of assembly gaps that have been identified: depth of coverage gaps, repetitive element-associated gaps, and muted gaps (see [6] for a recent review of gap types). Depth of coverage gaps occur when there are no sequence reads covering a particular genomic region. Typically, these gaps can be manually corrected by using PCR-based enrichment and sequencing. Repetitive element-associated gaps are observed when assembly algorithms cannot properly disambiguate repetitive sequences. Muted gaps are defined as genomic regions that are contiguous in an assembly, but which lack sequence found in many individuals in the species. There are two major sources of muted gaps: errors in bacterial cloning (when generating BACs) that delete sequence from assembly source DNA, and population variation representing true structural polymorphisms in an organism's genome [7].

In the following review, we will explore the strengths and weaknesses of various assembly algorithms as well as discuss how to evaluate and improve the quality of a genome assembly with a special focus on the human reference assembly. Finally, we will consider genome assembly techniques, challenges, and advances for nonhuman organisms.

II. ASSEMBLY ALGORITHMS

Starting with randomly sampled reads from the DNA, assembly algorithms aim to reconstruct the genomic sequence of an organism. As most vertebrate species are diploid (i.e., there exist two copies of each chromosome), an "optimal" assembler would generate two long strings for each chromosome; however, due to the fact that the two copies are highly similar to each other, algorithms historically reconstruct only one copy. This problem is even more pronounced in polyploid genomes where there are more than two copies of each chromosome [i.e., several plant genomes are tetraploid (four copies), hexaploid (six copies) or more]. Computationally, the assembly problem bears similarities to the shortest common superstring problem (SCS), which is known to be NP-complete [8], and it was previously proven that the assembly problem is NP-hard by reduction from SCS [9]–[11]. Given a set of n strings $\{s_1, s_2, \ldots, s_n\}$, and a length value L, the decision version of the

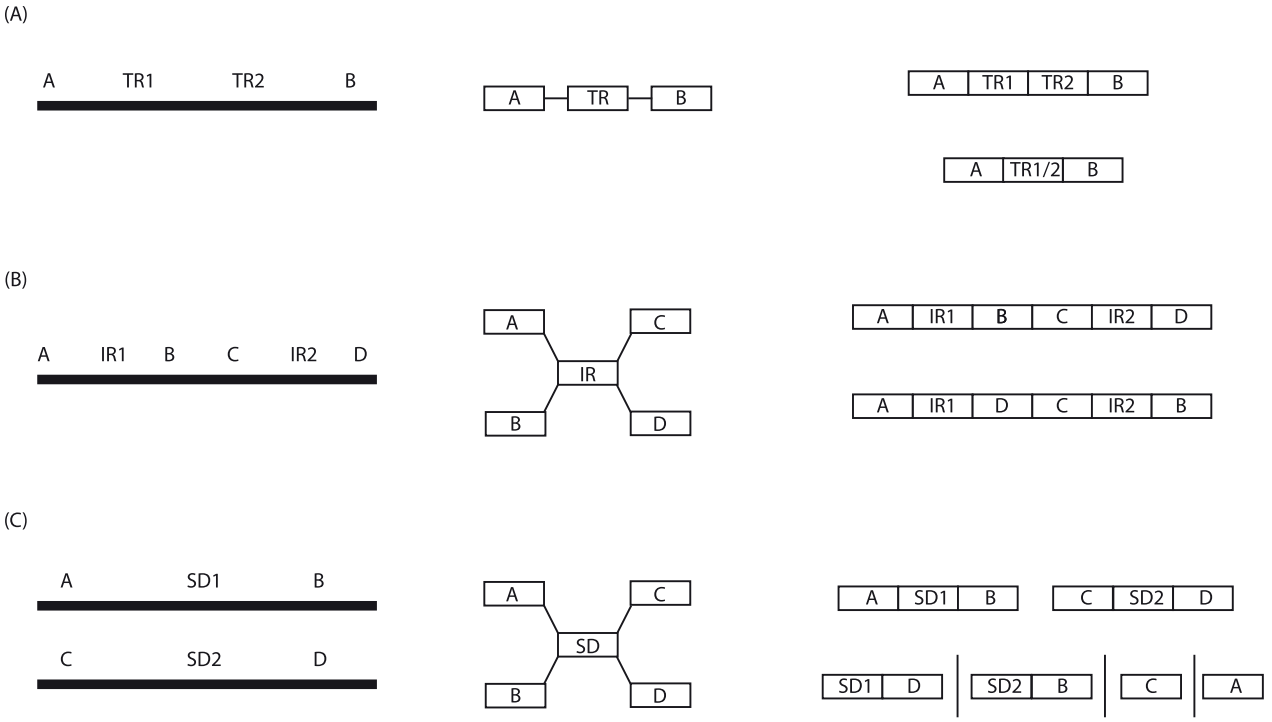

Fig. 1. ***The effects of repetitive elements on assembly representations. (a) Unique segments A and B are separated by a tandem repeat with two copies in the first panel; the assembly graph is delineated in the middle panel; the correct assembly is shown in the third panel on top and the collapsed assembly representation is shown below. (b) Unique segments A and B are separated by a single copy interspersed repeat, and unique segments C and D are separated by a single copy of the same repeat; the middle panel shows the assembly graph; the correct assembly is shown in the third panel on top and the inverted, incorrect assembly representation is shown below, where B and D are incorrectly assembled. (c) Unique segments A and B are separated by a segmental duplication, and unique segments C and D are separated by a highly identical copy of that segmental duplication; the graph is shown in the middle panel; this correct assembly representation is shown in the third panel on top, while the incorrectly assembled genomic segments are shown on the bottom. The misassembly creates gaps (vertical lines) in the assembly.***

SCS problem asks whether there exists a superstring S that contains each $s_i \leq 1 \leq i \leq n$) as a substring and $|S| \leq L$. Genome assembly problem differs from the SCS as follows.

1) The shortest common superstring problem tries to minimize the length of the superstring, where the genome assembly problem tries to achieve a superstring of a given length (i.e., $|S| \cong L$, where L is the genome size).
2) In mammals, approximately half of the genome is repeated, and plants have even more repetitive content, therefore two strings ($s_i s_j$, where $s_i = s_j$ and $i \neq j$) may be different substrings of the assembly.
3) Due to both sequencing errors and heterozygosity (where alleles at a genomic locus are different), a string *si* may include several substitutions, insertions, and deletions when aligned to its respective location within the superstring.
4) A string (approximately half of the input strings) may be present in the superstring in reversed and complemented form due to the double-helix structure of DNA.

The computational complexity of optimally solving the genome assembly problem led to the development of various heuristics, or reduction of the assembly problem to other NP-complete problems such as Hamiltonian path or traveling salesman problem, and simplification through several assumptions in such a way that lends itself to approximate solutions. Briefly there are three paradigms for genome assembly: 1) greedy construction of contigs; 2) overlap–layout–consensus (OLC); and 3) de Bruijn graph assembly (also reviewed in [12] and Table 2).

Greedy contig construction methods such as TIGR [13] and phrap [14], [15] start with computing prefix–suffix matches between all pairs of sequences. Next, they pick each sequence pair greedily, i.e., the two sequences with the "best" prefix–suffix match, as defined by match/alignment length and alignment sequence identity, are merged into a longer sequence. They then continue with the next pair showing the highest alignment score. Greedy assemblers may explicitly use graphs or a simple list as the main data structure to keep fragment overlaps. Note that some sequence pairs might be entirely identical, or one may subsume the other, which causes "collapsing" of such sequences in the final assembly. Therefore, greedy contig construction methods work well for genomes characterized by low or

Table 2 Assembly Algorithms

Algorithm type	Most common uses	Algorithm name	Reference
Greedy	Low repeat content genomes	TIGR	(Sutton et al. 1995)
		phrap	(Green 1994; Bastide and McCombie 2007)
Overlap-Layout-Consensus (OLP)	Low error, medium length sequence reads. Can be used with PacBio using pre- and post-assembly correction	Celera	(Venter et al. 2001)
		PCAP	(Huang et al. 2003)
		ARACHNE	(Batzoglou et al. 2002)
		Phusion	(Mullikin and Ning 2003)
de Bruijn graph	Short read sequences	EULER-SR	(Chaisson and Pevzner 2008)
		Velvet	(Zerbino and Birney 2008)
		ALLPATHS-LG	(Gnerre et al. 2011)
		SOAPdenovo	(Luo et al. 2012)
		Cortex	(Iqbal et al. 2012)
		SPAdes	(Bankevich et al. 2012)
		ABySS	(Simpson et al. 2009)
		Meraculous	(Chapman et al. 2011)
		HipMer	(Georganas et al. 2015)
String graph	Combine OLC and de Bruijn methods	SGA	(Simpson and Durbin 2012)
		Falcon	*
Hybrid	Use data from multiple platforms	Cerulean	(Deshpande et al. 2013)
		hybridSPAdes	(Antipov et al. 2015)
		Ray	(Boisvert, Laviolette, and Corbeil 2010)

*https://github.com/PacificBiosciences/falcon

no repeat content (i.e., some bacterial and viral genomes), however, they produce substantially shortened assemblies for genomes with repeats.

OLC-based assemblers were first successfully used to assemble the genome of the fruit fly (*D. melanogaster*) [16]. Although they work best with low-error medium length (700–1000 bp) sequence reads generated using the now mostly abandoned Sanger technology, applications for long reads generated by PacBio or similar platforms also exist. As the name implies, OLC methods are composed of three main steps. First, the prefix–suffix overlapsof all pairs of reads are calculated. The overlap step also creates an overlap graph where nodes represent reads and there are directed edges between reads that have sufficient-length prefix–suffix matches. Some OLC algorithms may also represent the length and identity of the matches as edge weights. Depending on the graph construction, the layout step then approximately calculates either Hamiltonian path in unweighted directed graphs or traveling salesman path in directed graphs. Because of the oversampling of DNA (i.e., depth of coverage >1), several reads may be represented within the same section of the layout, and each read in the same section may be slightly different from each other due to different sequencing errors. Therefore, the last step (consensus) calculates multiple sequence alignment (MSA) of the overlapping reads to determine the consensus sequence. Note that MSA problem is also NP-hard [17]. Some of the most popular OLC-based assemblers are Celera [18], PCAP [19], ARACHNE [20], and Phusion [21].

Because of the increased difficulty of all pairs alignments with PacBio and Oxford Nanopore (ONP) data due to higher sequencing errors, the OLC-based assemblers for such data either apply: 1) a preassembly error correction using orthogonal data from low error platforms (PacbioToCA [22]); 2) consensus calling through multiple sequence alignment (HGAP [23] and Nanocorrect [24]); or 3) efficient filters to reduce computational burden of all pairs alignments (MHAP [25]). Assembling high error reads is usually followed with postassembly error correction (or "polishing") using additional tools such as Quiver [23] or Nanopolish [24].

A secondary class of OLC-based assemblers use string graphs [18] The assembly problem on string graphs is formulated as minimum cost network flow problem [26]. Examples of string graph based assemblers include SGA [27] that uses Illumina, and Falcon (https://github.com/PacificBiosciences/falcon) that uses PacBio data. Recently, minimap and miniasm were developed to perform mapping and *de novo* assembly for PacBio SMRT and ONT reads without error correction [28].

For short-read assembly, the best option involves a de Bruijn-graph-based assembler, where the data structures and the algorithms are tightly coupled with each other. de Bruijn graphs are k-dimensional directed graphs on m symbols, where edges represent a size k-1 overlaps between strings generated by an m-symbol alphabet. In genome assembly problem, $m = 4$ (i.e., $\Sigma = \{A, C, G, T\}$) and k is the length of k-mers extracted from sequence reads. Theoretically, there can be 4^k nodes in a de Bruijn graph for genome assembly, although in practice, the genome length acts as an upper bound for node number. k is usually determined empirically by assembling the genome using different values, however, preassembly estimation is also possible by analyzing k-mer abundance in the raw sequence data [29]. As a preprocessing step, it is common to remove k-mers with very low frequency. de Bruijn-based assemblers represent each k-mer as a node and length k-1 overlaps as edges in the graph. Note that both in- and out-degrees of each node are at most 4,

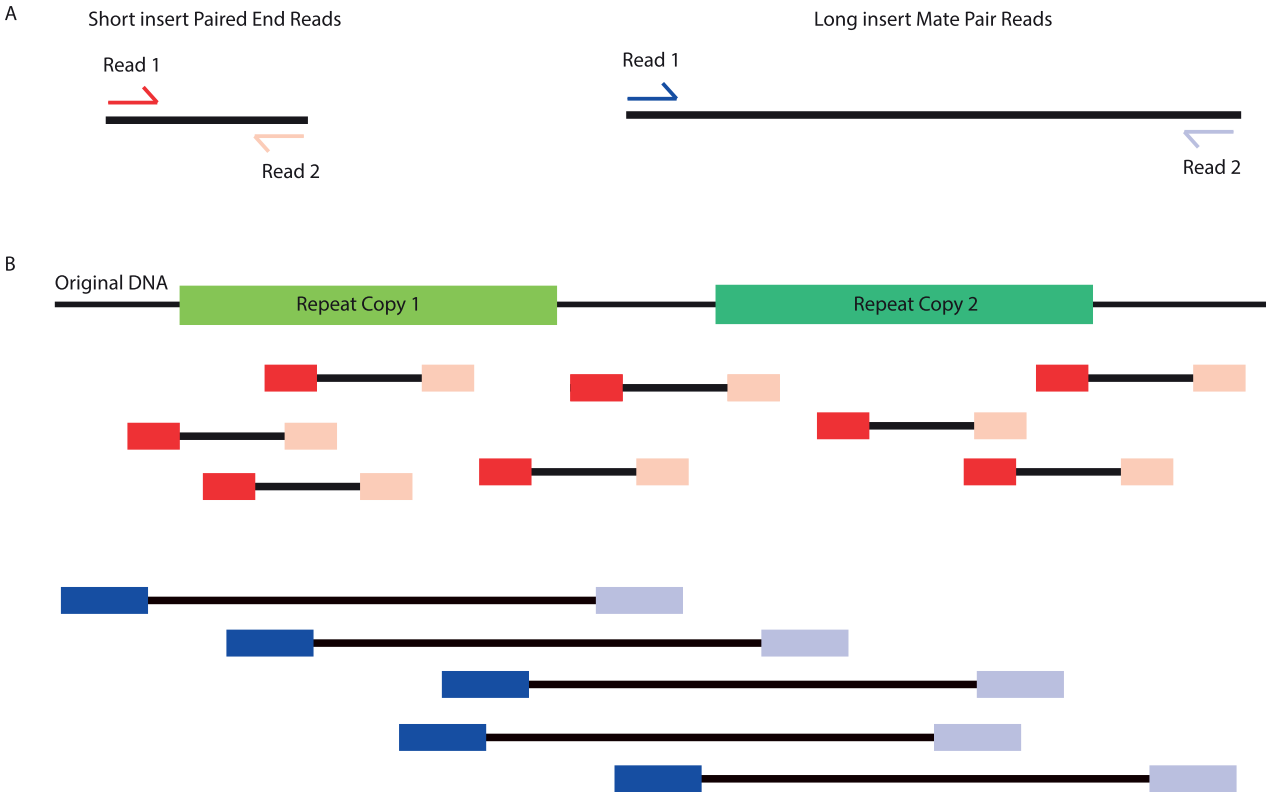

Fig. 2. ***Paired ends and mate pairs. (a) The major difference between paired ends and mate pairs is the length of the insert. Paired ends are generally short (100–500 bp) while mate pairs are longer (2–5 kb). (b) During de novo assembly, the large inserts (mate pairs) are able to correctly assemble in complex regions of repetitive elements (green boxes) while short inserts (paired ends) can fill in gaps missed by mate pairs.***

and all-pairs alignments are avoided by simply linking *k*-mers that are adjacent in any read during graph construction. Sequencing errors present themselves as either bubbles (i.e., two outgoing paths from a node are later merged on the same node) or tips (i.e., dead-end paths) in the graph, which are resolved using approaches that vary among different assemblers. Next, the assembler constructs the contigs by simply finding Eulerian paths [30], [31]. Note that Eulerian path represents sequence since the graph construction also ensures all edges to represent a sequence of length *k-1*. The use of de Bruijn graphs for genome assembly was first proposed by Pevzner for sequencing by hybridization [32]. Later, with the development of capillary Sanger sequencing, Pevzner *et al.* [30], [31] again formulated a novel de Bruijn-based genome assembly algorithm using these read data. EULER-SR [33] was the first algorithm to use de Bruijn graph assemblies on Illumina/Sanger hybrid data, which was followed with pure Illumina assembly by Velvet [34], and incorporating mate-pair sequencing by EULER-USR [35]. Mate pairs are pairs of oppositely oriented sequence reads from a single fragment of DNA. Mate-pair libraries are generated with longer insert sizes (2–5 kbp). Paired ends are also pairs of reads from a single fragment of DNA; however standard paired end libraries are generally much smaller (200–500 bp) (see Fig. 2). Other popular de Bruijn-based assemblers include ALLPATHS-LG [36], SOAPdenovo [37], and Cortex [38]. Recently, paired de Bruijn graphs [39] and pathset graphs [40] were developed to directly represent mate-pair information within the assembly graph, as implemented by the SPAdes assembler [41]. The computational burden of genome assembly also spearheaded development of distributed approaches such as ABySS [42], Ray [43], and Meraculous [44] use message passing interface (MPI) in commodity clusters, and more lately, HipMer [45] Unified Parallel C in extreme scale supercomputers.

Although there are several tools that use OLC or string graphs to assemble reads generated with the single molecule read technologies (i.e., PacBio, ONP) as outlined above, their relatively higher cost for data generation also prompted the development of hybrid assembly algorithms. These algorithms can either simultaneously use data generated using multiple platforms (Cerulean [46], hybridSPAdes [47]), or make use of the data from different platforms one by one, in a hierarchical fashion (Ray [43]). Such methods may use either of the assembly strategies outlined above, and some employ several of these techniques to obtain a draft assembly. The SPAdes package also includes additional algorithms (truSPAdes [48]) that use "long virtual reads" generated with the Illumina TruSeq long-read technology.

The assembly algorithms first generate contiguous strings of DNA (contigs) that contain no gaps. Typically, depending on the genome, read, and fragment lengths, the initial assembly comprises thousands to hundreds of thousands, and in some cases, millions of contigs. These contigs are later ordered and oriented with respect to one another through a postprocessing step called scaffolding. The scaffolders typically use paired-end, mate-pair, or long-read information to "link" contigs to form longer scaffolds. While many of the assembly tools have built-in scaffolders, there are also standalone scaffolding algorithms, such as SSPACE [49], SCARPA [50], BESST [51], and OPERA [52]. A more recent tool, LINKS [53], can use ONP reads to build scaffolds. Some scaffolding algorithms use external data sources (i.e., data other than the original input sequences), such as optical mapping (Hybrid Scaffold [54]), or RNA-seq data (RNAPATH [55], [56]). Finally, the gaps between contigs within scaffolds are filled in using either built-in (e.g., SOAPdenovo Gap Closer [57]) or standalone (e.g., Sealer [58]) gap closing algorithms.

III. ASSESSING ASSEMBLY QUALITY

Genome assembly is computationally complex, and the output is typically a haploid representation of the diploid genome, reconstructed from the given reads. For this reason, it is essential to assess the quality of an assembly by examining errors at both the single nucleotide level as well as at the level of a potential large-scale misassembly. Statistics that assess assembly contiguity include N50, L50, and NG50 lengths (see Table 3) [59], [60]. Mate-pair information and depth of coverage are additional metrics that can indicate assembly errors. The known distribution of distances between the two pairs and their orientation can be used to validate assembly quality. Mate pairs can also be used to resolve repeats during the assembly process (see Fig. 2). Based on mate-pair library construction techniques, we can expect their insert size, as well as order and orientation, to be consistent across the genome. When any of these intrinsic assembly measures are inconsistent, tools such as AMOSvalidate [61], Tool for Analyzing mate pairs in Assembly (TAMPA) [62] and Recognition of Errors in Assemblies using Paired Reads (REAPR) [63] are recommended for flagging misassembled regions for manual review. For example, excess depth of coverage in a region can be indicative of erroneously collapsed repetitive elements. Likewise,

Table 3 Assembly Statistics

Term	Definition
N50	50% of all bases in entire assembly are contained in contigs equal to or larger than the value, N
L50	Number of contigs that equals N50 when their lengths are summed
NG50	50% of the estimated genome size is equal to or larger than the value, N.

Table 4 BioNano Map and PacBio Assembly Combined to Create Hybrid Scaffold of a Hydatidiform Mole, CHM13

	BioNano Map	PacBio Assembly	Hybrid Scaffold
# of Contigs	3593	1590 *	254
Min Contig Length	0.08 Mb	0	0.27 Mb
Median Contig Length	0.61 Mb	0.06 Mb	4.35 Mb
Mean Contig Length	0.78 Mb	1.78 Mb	9.68 Mb
Contig N50	**1.02 Mb**	**13.46 Mb**	**20.79 Mb**
Max Contig Length	5.27 Mb	63.15 Mb	82.83 Mb
Total Contig Length	2.812 Gb	2.824 Gb	2.458 Gb

* Number of contigs used in hybrid scaffolding

an overabundance of mismatches in overlapping sequence reads comprising a contig is a common hallmark of assembly error. Yet, these error detection tools often have high false positive rates that can be attributed to inaccuracy and noisiness in experimental protocols that can lead to nonrandom coverage across the genome and mispaired reads [61].

Independent data from complementary technologies can also be used for assessing and improving assemblies. It is often helpful to evaluate assemblies with sequence that did not contribute to the assembly, especially if such sequence has been generated with a different sequencing technology. These sequences provide an autonomous measure for evaluation of potential sequence or assembly error [64]. Additionally, whole genome or optical mapping is a high-resolution single molecule technique for resolving assembly errors and improving assembly statistics. In brief, single molecules of high molecular weight DNA are elongated and passed through nanochannels while being nicked or cut at known restriction fragment length polymorphism sites and labeled with fluorescent markers [65]. These molecules are then stained, imaged, and measured in real time without any cloning, amplification, or hybridization steps. The resulting restriction maps are assembled into a whole genome consensus restriction map. This provides long-range (several hundred thousand base pairs or greater) genomic information, but at the costs of a higher base error rate and lower sequence resolution. For these reasons, whole-genome mapping is currently best used to complement other assembly approaches [66]. For example, we used a BioNano (http://www.bionanogenomics.com/) genomic map in conjunction with PacBio long-read data to improve a human assembly where the contig N50 length increased from 13.46 to 20.79 Mbp (Fig. 3 and Table 4).

The choice of assembly algorithm can also impact assembly quality and contiguity. Two major efforts to evaluate and benchmark assembly software were the Assemblathon [60], [67] and Genome Assembly Gold-Standard Evaluation (GAGE) [68]. The Assemblathon 1 used simulated genome reads to test the performance of various assembly algorithms, while the Assemblathon 2 used real sequence data, generated by a variety of technologies, from three different vertebrate species' genomes. Software developers submitted assemblies that were then scored based on contiguity, completeness, and accuracy. The GAGE

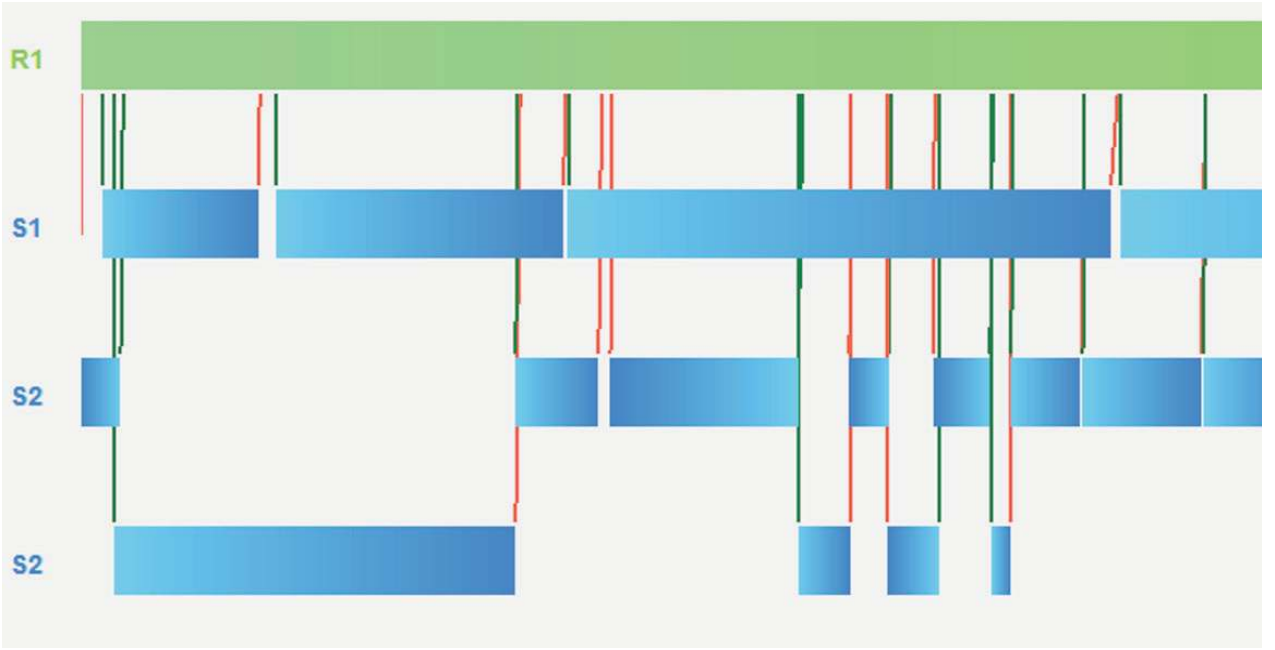

Fig. 3. ***PacBio and BioNano hybrid scaffold. The green box (R1) is the 100% contiguous hybrid scaffold created by combining the BioNano genomic map (blue boxes labeled S1) data with PacBio long reads (blue boxes labeled S2). Gaps in either technology are filled by the complementary technology to increase scaffold N50.***

competition used Illumina read-based assemblies to assess four different algorithms. Both efforts highlighted how different pipelines and protocols produced variable assemblies. Whereas GAGE concluded that input data quality was the most important factor in assembly quality, the Assemblathon posited that each data set should be assembled using multiple algorithms due to the lack of an optimal assembly pipeline. No single pipeline produced an assembly that excelled at all metrics assessed.

Quality control of assemblies can be achieved by comparing two assemblies using assembly–assembly alignment software, such as the BLAST-based algorithm developed at NCBI (https://www.ncbi.nlm.nih.gov/genome/tools/remap) [69]. For example, a human assembly can be aligned to the human reference genome assembly and the proportion of the reference coverage can be calculated to indicate the input assembly's quality. NCBI defines the alignments from its assembly comparison process as either first pass or second pass. First pass alignments essentially embody reciprocal unique alignments between the query and target assemblies. The first pass alignments represent the output of an alignment merge and clean-up algorithm in which BLAST results are processed to identify genomic regions containing components common to both assemblies, that are then merged into a common component set, defined as the longest, most consistent stretches of sequences generated from alignment. Redundant alignments are removed and conflicting alignments are evaluated in the second pass. Second pass alignments are defined for large (>1 kbp) regions that have no conflicting alignments in the first pass. Query loci in second pass alignments can map to multiple locations in the target assembly. The second pass often helps resolve duplicated sequences that have been expanded or collapsed in the query assembly relative to the target when the reference assembly is the target assembly. Alternatively, QUAST [70] or ALE [71] can be used for assembly quality assessment which calculates several metrics of a given assembly with or without a "guide" reference genome. QUAST and ALE are best used for microbial genome assemblies in their current implementation and are not ideal for large eukaryotic genomes. Similar tools such as LAP [72] and CGAL [73] demonstrate support for eukaryotic genomes. All of these tools are based upon the idea that the assembly must be consistent with the data generation process first introduced by Myers in 1995 [74].

Another method for assessing assembly quality is to assess gene content by aligning within-species or closely related RefSeq transcripts to the assembly. The Reference Sequence (RefSeq) collection is a comprehensive, nonredundant set of sequences that are annotated and curated by NCBI [75]–[77]. They include genomic DNA, transcripts, and proteins. By comparing alignments of these data to various assemblies from the same or related species, it is possible to calculate the percentage of genes represented within an assembly [78]. Data concerning the number of transcripts that are complete versus those that are split, or those where the alignment is low quality serve as a proxy for assessing assembly continuity and correctness. Analyses of frameshifts, which are disruptions to transcript reading frames that are expected to alter

protein function, in such alignments can also be used to assess assembly quality [79].

IV. HUMAN REFERENCE ASSEMBLY

In 2001, the Human Genome Project (HGP) [80] released a draft sequence of the human genome, an achievement hailed as a turning point in human genetics. In a parallel effort, the private company Celera used whole genome shotgun methods with Sanger-based sequencing technology to produce an additional human genome assembly [81]. Representing 90% of euchromatic sequence, the HGP draft assembly was superseded in 2004 by the "finished" human reference assembly, representing 99% of the euchromatic sequence, and accurate to an error rate of 1 in 100 000 [82]. Ongoing curation by the Genome Reference Consortium (GRC; https://www.genomereference.org) continues to improve the reference by resolving remaining issues such as gaps or misassemblies, and adding new sequence representations that capture human population diversity. The human reference genome represents the highest quality mammalian genome ever assembled, due in part to the sequencing and assembly approaches that were used.

In contrast to the whole genome shotgun (WGS) assembly approach that is used for most assemblies derived from next generation sequencing reads, the human reference assembly was constructed using a clone/map-based approach, using Sanger-sequenced genomic clones (see description of clone-based assemblies above). Relative clone order and a minimal tiling path were determined by a combination of genetic, radiation hybrid (RH), fluorescent *in situ* hybridization (FISH) maps [83] and fingerprint maps [84]. Selected clones from the tiling path were then fragmented, subcloned, and shotgun sequenced. These fragments were then shotgun assembled to recreate the consensus sequence of each clone using software called phrap [14]. Clones were manually finished with polymerase chain reaction (PCR) to close gaps resulting from the shotgun approach. The genome sequence was then assembled by aligning overlapping clones to create a minimal tiling path of nonredundant sequence.

Notably, the clone-based assembly approach differs from the WGS approach because it reduces the *de novo* assembly problem from a whole-genome level to a local, clone-based one [85], [86]. As a result, clone-based assemblies have substantially longer N50s and fewer gaps than WGS assemblies sequenced with corresponding methods and generally exhibit less collapse of repetitive or segmentally duplicated sequence than WGS assemblies [2], [7]. In contrast to WGS assemblies, which typically provide a consensus haploid representation for diploid genomes, clone-based assemblies create a mosaic representation, with clone-boundaries serving as potential haplotype boundaries. A haplotype block is a set of linked markers on a chromosome that are inherited together. Furthermore, haplotype blocks are generally longer than those found in WGS assemblies, as each clone represents a single haplotype, and variations in clone overlaps can be used to identify and assemble a path of clones representing the same haplotype. However, it should be noted that a small number of genomic regions are recalcitrant to cloning and result in sequence gaps that must be resolved by complementary assembly approaches. In addition, when clone sequences used for assembly represent highly divergent haplotypes, such as those associated with polymorphic structural variants, it can lead to assembly errors and creation of mixed haplotypes not observed in the population [87], [88]. Despite the high quality of assembly that can be achieved with the clone-based approach, this technique has nonetheless generally been abandoned in favor of the WGS approach, due to the high costs associated with cloning, sequencing, and mapping, as well as improvements in WGS assembly quality due to the availability of long reads and updated algorithms.

It should be emphasized that because current technologies do not permit sequencing of entire chromosomes, all assemblies, clone based or WGS, are genome models, not actual genomes. Although a haploid chromosome can be represented as a linear chromosome sequence, the assembly of such a chromosome from diploid source DNA can be confounded in regions of significant haplotypic diversity. This is further compounded in the human reference genome, which is derived from DNA representing many diploid individuals [80]. The linear haploid chromosome model initially used by the HGP for the human reference genome assembly did not have a robust mechanism for representing the diversity of such regions. As noted previously, the coplacement of divergent haplotypes in a single linear sequence can lead to sequence representations not found in individuals or, if sufficiently divergent, assembly gaps [87]–[90]. An assembly model developed by the GRC improved the representation of divergent regions in the human reference genome. Retaining the linear chromosomes and unlocalized and unplaced scaffolds (sequences that cannot be localized to specific chromosome regions or chromosomes, respectively) of the original model, it introduced the concept of alternate loci scaffolds [91]. These scaffold sequences permit the representation of structurally complex regions as multiple independent sequence paths within the assembly. However, these alternate paths can be given chromosome context by virtue of their alignment to the corresponding chromosome path. This model is used for the current human and mouse reference genome assemblies (Fig. 4).

In addition to providing a mechanism for improved representation of complex assembly regions, the alternate loci permit the reference assembly to include multiple sequence representation for any region. As a result, this assembly model supports the representation of population diversity and assemblies using this model cannot be considered haploid genome representations. The current major release of the human reference genome assembly, GRCh38 (GCA_000001405.15), contains 261 alternate loci scaffold sequences representing 178 nonoverlapping genomic regions containing over 150 genes not on the primary assembly

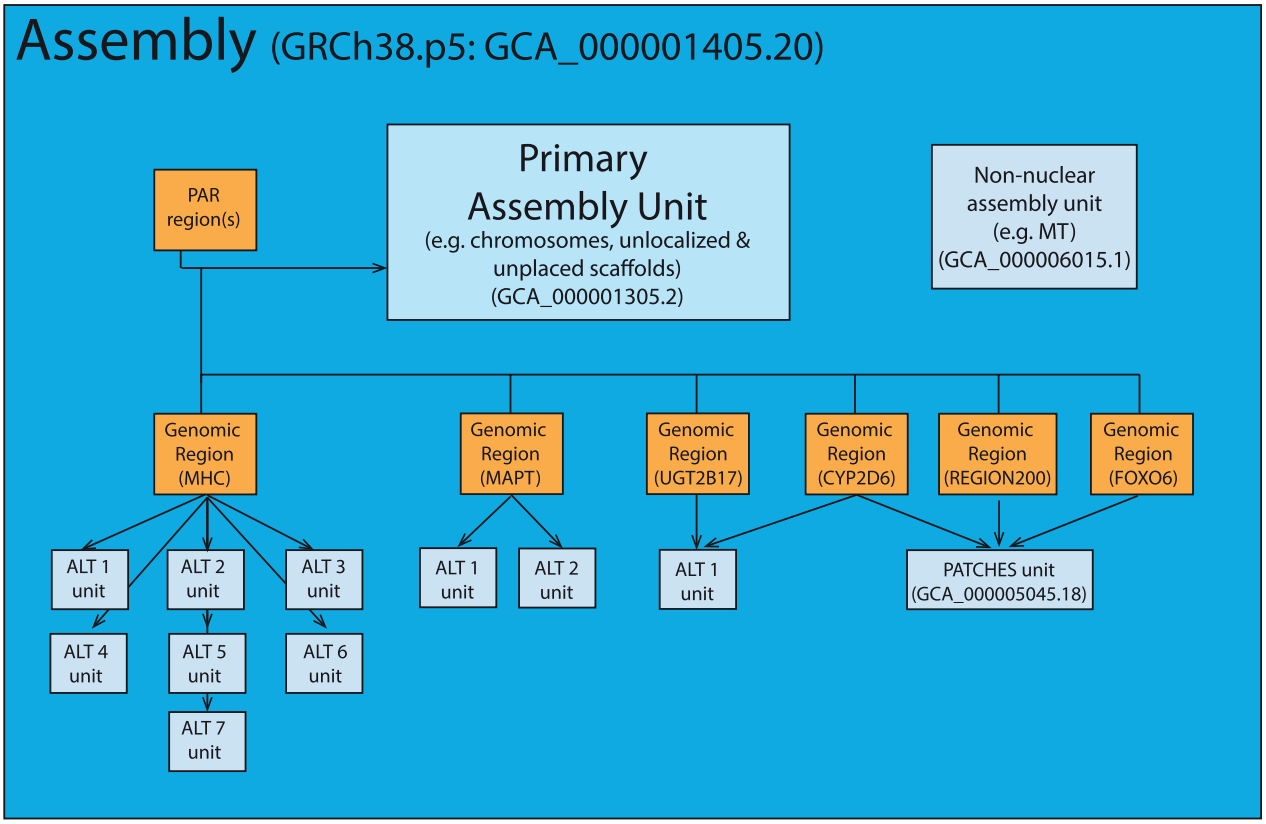

Fig. 4. ***Graphical representation of the NCBI assembly model (https://www.ncbi.nlm.nih.gov/assembly/model/). Accessions and regions shown correspond to the GRCh38.p5 version of the human reference genome assembly. The assembly model used by the HGP was extended to account for sequences that do not fit in the linear chromosome space. Genomic regions are defined on the primary assembly unit. Alternate loci and patch scaffolds provide alternate sequence representations for the genomic regions and are assigned to discrete assembly units. The first alternate locus scaffold representing each region is assigned to the same alternate loci assembly unit (e.g., ALT 1 unit). Patches and alternate loci scaffolds are given chromosome context via alignment to sequences in the primary assembly unit wherever possible. The full assembly and each assembly unit is assigned a versioned accession.***

(ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA_ 000001405.15_GRCh38) [92].

The human reference genome assembly is unusual among genome assemblies, in that it is the subject of ongoing curation efforts. This work is performed by the GRC, which also curates the mouse, chicken, and zebrafish reference assemblies. Improvements found in GRCh38 can be broadly categorized as sequence additions, sequence corrections, and tiling path updates to chromosomes and increased representation of diversity in the form of new alternate loci. Various assembly resources and techniques were involved in implementing these changes [93]. Paired-end mapping of large insert genomic clones to the reference was used to identify inversions, sequences that extend into or span assembly gaps, misassemblies and regions that are candidates for alternate sequence representation [94]. Clones mapping to regions of interest were sequenced in entirety, and added to chromosomes or alternate loci scaffolds.

Clones from a BAC library derived from an essentially haploid hydatidiform mole DNA source, CHM1 [65], [95], played an important role in the detection and resolution of several reference assembly errors in the latest assembly update. A hydatidiform mole is an aberrant form of pregnancy where an enucleated egg is fertilized with a sperm, the sperm DNA doubles, and the cells grow unchecked. This tissue material is valuable as it contains only paternal germline DNA and is a haploid representation of the human genome. Having only one haplotype (as opposed to two for a diploid human genome) makes assembly through repetitive regions and complex structural architecture much easier. In addition to their use in the paired-end mapping analyses, CHM1 BAC clones were used to generate single haplotype assemblies of regions previously misassembled due to haplotype mixing from multiple DNA sources such as the immunoglobulin heavy chain variable locus [96]. These local assemblies were then integrated into the GRCh38 assembly. Other CHM1 BAC paths were added as alternate loci scaffolds to represent additional variation in the assembly.

GRCh38 is considered an improvement over prior reference assembly versions by various metrics [93]. Contig and scaffold N50s both increased, as did the ungapped sequence length and the count of unspanned gaps (https://www.ncbi.

nlm.nih.gov/assembly/GCF_000001405.13, https://www.ncbi.nlm.nih.gov/assembly/GCF_000001405.26). Furthermore, the current assembly exhibits improved annotation statistics, when compared to GRCh37 (https://www.ncbi.nlm.nih.gov/genome/annotation_euk/ Homo_sapiens/106/, https://www.ncbi.nlm.nih.gov/genome/ annotation_euk/Homo_sapiens/105/). Specifically, there is an increase in the number of annotated genes, and a decrease in genes whose annotation is interrupted by assembly gaps (split genes) or low consensus coding sequence (CDS) coverage. GRC efforts to improve the human reference assembly are ongoing. The human reference assembly is assigned a unique, versioned assembly accession by GenBank, an INSDC database (GRCh38: GCA_000001405.1). Sequence updates to the assembly increment the accession version and all changes are tracked in the NCBI assembly database [97]. In addition to the assembly accession, every assembly sequence has its own versioned accession, which also increments when there are sequence changes.

In addition to major assembly releases, in which chromosomes and/or other sequences in the primary assembly unit are updated, the GRC assembly model supports minor releases, known as patch releases, that provide users with timely access to assembly updates without disruption to chromosome coordinates [91]. Patch releases include sequence corrections (fix patches) and new sequence representations (novel patches). The former prefigure chromosome updates in the next major assembly release, while the latter are operationally equivalent to alternate loci scaffolds. Patch releases increment the assembly accession version, but not chromosome accession versions, because the patches exist as scaffolds and do not create changes to chromosome sequences or coordinates. In a major release, fix patch scaffolds are incorporated into chromosomes, novel patch scaffolds are redefined as alternate loci, and both assembly and sequence accession versions are updated. Patch releases of the human reference assembly occur regularly, to provide rapid access to sequence updates, while major assembly releases occur infrequently in order to minimize reassignment of annotations to new sequence coordinates.

V. NONHUMAN GENOME ASSEMBLY

The number of nonhuman eukaryotic draft quality assemblies generated from short (100 bp) Illumina reads has grown tremendously over the last decade, and numerous laboratories around the world have made substantial progress in both creating and improving whole genome reference assemblies from these billions of short sequences. Creating the first draft assemblies of nonhuman species with next-generation sequencing technology enabled researchers to begin initial genomic analyses while balancing budgetary constraints associated with genome assembly. In cases where whole genome models (reference assemblies) were unavailable, these draft assemblies were typically generated *de novo* using a variety of de Bruijn-graph-based assembly algorithms. Laboratories with expertise in genome assembly have routinely produced genome assemblies using this algorithm for protists, birds, fish, mammals, insects, and more. The most popular assembly algorithms, ALLPATHS-LG and SOAPdenovo2, dictate ~100 × total sequence coverage of tiered insert length paired or overlapping sequences [36], [57].

Upon completing a nonhuman genome assembly, a primary goal is to annotate for gene content, while also improving the accuracy of the sequence assembly as a byproduct of transcriptome sequence data. Following these steps, downstream genetic experiments typically commence in order to dissect phenotypes that oftentimes have implications for human disease. Usually, with the type of assembly described above, the contig lengths will be sufficient for gene predictions and postassembly alignment-based analysis. However, many partial gene models may remain and often can thwart further analysis, i.e., orthology.

Furthermore, it is clear that the limited contiguities of most initial draft assemblies are not entirely sufficient for detailed genetic investigations of the molecular signatures of selection or for modeling putative disease causing alleles. Moreover, while the diversity of species with genome assemblies generated for the purpose of further experimentation rapidly grows, more complete representations of nonhuman draft assemblies remain woefully behind. Nonetheless investigators can still use incomplete genome draft assemblies to add critical knowledge to our growing understanding of biological systems. As a result of this preliminary success, there is a growing broader interest in raising the assembly quality standards for several of the more heavily used nonhuman genome models. This interest, in turn, has compelled a renewed emphasis on assembly quality across all levels of biology.

At the same time, many propose to instantiate better nonhuman genome assemblies by taking advantage of new sequencing and assembly developments by transitioning to the use of long, single-molecule sequences, specifically PacBio technology, combined with high-resolution optical maps that can accurately resolve the majority of gross assembly errors, with the hope of obtaining near-complete copies of chromosomes [25], [98]. Such technologies can dramatically change the outcome of nonhuman genome assembly projects by capturing large swaths of sequence with fewer gaps, albeit at higher cost. Although quality metrics for some of the first assemblies derived from updated technologies and algorithms are encouraging for relatively higher measures of contiguity, many still require improved sequence representation, especially in regions with base-composition bias. As part of a recent proposal to improve genome assemblies for a group of taxa that serve as aquatic models for human diseases, the most current draft assemblies for the species targeted for improvement displayed a range of missing bases that was conservatively estimated to total 1 Gb each. An additional 2%–5% of the estimated genome size was either not sequenced or assembled outside of scaffolds, within gaps or unplaced regions.

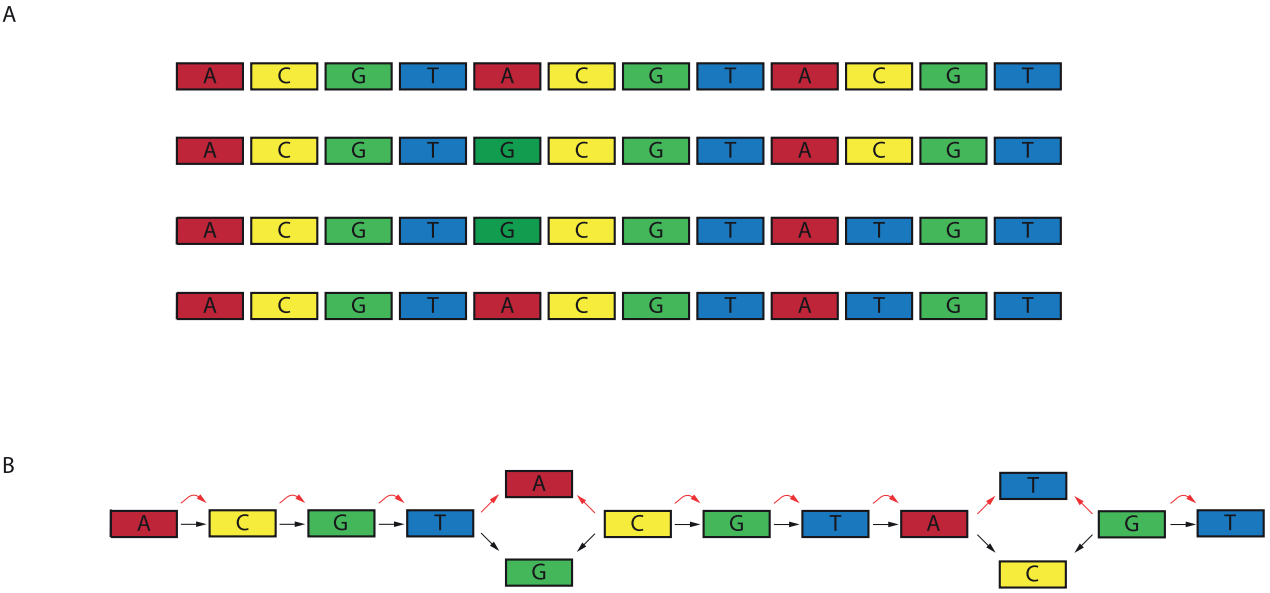

Fig. 5. ***Graph genome representation. (a) The linear chromosomes of four individuals. (b) The graph representation of the population, where polymorphisms are represented as "bubbles" in the graph.***

Recent commentaries address these issues surrounding *de novo* assembly [99], [100], and previous efforts to close gaps in these particular assemblies, as expected, revealed that many assembly gaps were in regions containing structurally variant alleles, simple tandem arrayed repeats, and high GC content.

VI. FUTURE DIRECTIONS

Most reference genome assemblies are represented as a linear sequence, but it is clear that a simple assembly model of a single haploid representation is insufficient for modeling most populations given the nature and extent of genetic variation. Ultimately, taking multiple genomes from the same (or closely related) species and constructing a "pan genome" that comprehensively catalogs all variation in a given population would be ideal. Unfortunately, most existing sequence analysis tools still expect a haploid assembly and cannot accurately handle a multi-allelic reference [101]. An optimal reference assembly should be represented by a graph in which shared sequences are depicted by nodes while population and individual specific sequences are depicted by edges and branch points in the graph (Fig. 5). Recently, various groups have proposed graph structures for representing the human reference assembly [91], [102] to account for the vast amount of diversity present in the human population. Additionally, Iqbal *et al.* [38] proposed the de Bruijn graph for pan genome analysis, and Marcus *et al.* [103] developed a novel algorithm, splitMEM, for constructing a compressed de Bruijn graph (where the nodes and edges are compressed whenever the path between nodes is nonbranching) from a generalized suffix tree of input genomes. The algorithm decomposes the minimal exact matches (MEMs) from the suffix tree and extracts overlapping components to compute the nodes.

Novel technologies that leverage short-read sequencing to improve assembly contiguity have recently emerged in the field of genome assembly. Although traditional short-read sequences have many limitations for genome assembly, they have been used to resolve the 3-D structure of chromosomes in living cells using methods like Hi-C [104] and chromatin capture [105], [106]. DNA segments that are physically in close proximity are likely to be ligated and thus sequenced together as pairs, and the long-range information is used to assist in scaffolding and haplotype phasing [107]. A novel method that does not rely on living cells, called Chicago (Cell-free Hi-C for Assembly and Genome Organization) was developed by Dovetail Genomics [108]. This method uses reconstituted chromatin to connect DNA segments up to several hundred kilobases and the data can be used for both genome assembly as well as haplotype phasing and identification of structural variants.

New technologies also have the potential to increase sequence read lengths. DNA sequencing with nanopores was proposed nearly 20 years ago [109], and nanopore sequencing recently gained traction in the realm of genome assembly. This advance was largely contributed by the development of the MinION from Oxford Nanopore, a USB memory stick sized nanopore sequencer that can sequence long stretches of DNA from a single molecule in minutes. It has been used to successfully and rapidly assemble bacterial genomes [24] as well for detecting a hospital outbreak of *Salmonella enterica* in real time [110]. While the technology is promising, there are still limitations with respect to base quality and translational applications using more complex eukaryotic genomes.

In conclusion, the field of genome assembly is constantly and rapidly changing. New technologies continue to improve, allowing for longer reads that traverse the most complex regions of the genome characterized by repetitive elements and structural variants. ■

Acknowledgement

The authors would like to thank two anonymous reviewers for careful reading and excellent suggestions. They would also like to thank S. Kiwala, A. Coffman, J. Eldred, and D. Larson for reading the manuscript and providing comments.

REFERENCES

[1] E. S. Lander and M. S. Waterman, "Genomic mapping by fingerprinting random clones: A mathematical analysis," *Genomics*, vol. 2, no. 3, pp. 231–239, Apr. 1988.

[2] C. Alkan, S. Sajjadian, and E. E. Eichler, "Limitations of next-generation genome sequence assembly," *Nature Methods*, vol. 8, no. 1, pp. 61–65, 2011.

[3] J. A. Bailey *et al.*, "Recent segmental duplications in the human genome," *Science*, vol. 297, no. 5583, pp. 1003–1007, Aug. 2002.

[4] The Arabidopsis Genome Initiative, "Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*," *Nature*, vol. 408, no. 6814, pp. 796–815, Dec. 2000.

[5] P. S. Schnable *et al.*, "The B73 maize genome: Complexity, diversity, and dynamics," *Science*, vol. 326, no. 5956, pp. 1112–1115, Nov. 2009.

[6] M. J. P. Chaisson, R. K. Wilson, and E. E. Eichler, "Genetic variation and the *de novo* assembly of human genomes," *Nature Rev. Genet.*, vol. 16, no. 11, pp. 627–640, Nov. 2015.

[7] E. E. Eichler, R. A. Clark, and X. She, "An assessment of the sequence gaps: Unfinished business in a finished human genome," *Nature Rev. Gen.*, vol. 5, no. 5, pp. 345–354, May 2004.

[8] K.-J. Rälhä and E. Ukkonen, "The shortest common supersequence problem over binary alphabet is NP-complete," *Theor. Comput. Sci.*, vol. 16, no. 2, pp. 187–198, 1981.

[9] N. Nagarajan and M. Pop, "Parametric complexity of sequence assembly: Theory and applications to next generation sequencing," *J. Comput. Biol.*, vol. 16, no. 7, pp. 897–908, Jul. 2009.

[10] J. Kececioglu and E. Myers, "Exact and approximate algorithms for the sequence reconstruction problem," *Algorithmica*, 1995.

[11] P. Medvedev, K. Georgiou, G. Myers, and M. Brudno, "Computability of models for sequence assembly," in *Algorithms in Bioinformatics*, vol. 4645, R. Giancarlo and S. Hannenhalli, Eds. Berlin, Germany: Springer-Verlag, 2007, pp. 289–301.

[12] N. Nagarajan and M. Pop, "Sequence assembly demystified," *Nature Rev. Gen.*, vol. 14, no. 3, pp. 157–167, Mar. 2013.

[13] G. G. Sutton, O. White, M. D. Adams, and A. R. Kerlavage, "TIGR assembler: A new tool for assembling large shotgun sequencing projects," *Genome Sci. Technol.*, vol. 1, no. 1, pp. 9–19, Nov. 1995.

[14] P. Green, *Phrap*, vol. 36. 1994. [Online]. Available: http://www.genome.washington.edu/UWGC/analysistools/phrap.htm

[15] M. de la Bastide and W. R. McCombie, "Assembling genomic DNA sequences with PHRAP," *Current Protocols Bioinf.*, pp. 11–14, 2007.

[16] M. D. Adams *et al.*, "The genome sequence of Drosophila melanogaster," *Science*, vol. 287, no. 5461, pp. 2185–2195, Mar. 2000.

[17] W. Just, "Computational complexity of multiple sequence alignment with SP-score," *J. Comput. Biol.*, vol. 8, no. 6, pp. 615–623, Jul. 2001.

[18] E. W. Myers, "The fragment assembly string graph," *Bioinformatics*, vol. 21, no. 2, pp. ii79–ii85, Sep. 2005.

[19] X. Huang, J. Wang, S. Aluru, S.-P. Yang, and L. Hillier, "PCAP: A whole-genome assembly program," *Genome Res.*, vol. 13, no. 9, pp. 2164–2170, Sep. 2003.

[20] S. Batzoglou *et al.*, "ARACHNE: A whole-genome shotgun assembler," *Genome Res.*, vol. 12, no. 1, pp. 177–189, Jan. 2002.

[21] J. C. Mullikin and Z. Ning, "The phusion assembler," *Genome Res.*, vol. 13, no. 1, pp. 81–90, Jan. 2003.

[22] S. Koren *et al.*, "Hybrid error correction and de novo assembly of single-molecule sequencing reads," *Nature Biotechnol.*, vol. 30, no. 7, pp. 693–700, Jul. 2012.

[23] C.-S. Chin *et al.*, "Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data," *Nature Methods*, vol. 10, no. 6, pp. 563–569, Jun. 2013.

[24] N. J. Loman, J. Quick, and J. T. Simpson, "A complete bacterial genome assembled *de novo* using only nanopore sequencing data," *Nature Methods*, vol. 12, no. 8, pp. 733–735, Aug. 2015.

[25] K. Berlin, S. Koren, C.-S. Chin, J. P. Drake, J. M. Landolin, and A. M. Phillippy, "Assembling large genomes with single-molecule sequencing and locality-sensitive hashing," *Nature Biotechnol.*, vol. 33, no. 6, pp. 623–630, Jun. 2015.

[26] R. K. Ahuja, T. L. Magnanti, and J. B. Orlin, *Network Flows: Theory, Algorithms, and Applications*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1993.

[27] J. T. Simpson and R. Durbin, "Efficient *de novo* assembly of large genomes using compressed data structures," *Genome Res.*, vol. 22, no. 3, pp. 549–556, Mar. 2012.

[28] H. Li, "Minimap and miniasm: Fast mapping and *de novo* assembly for noisy long sequences," *Bioinformatics*, vol. 32, no. 14, pp. 2103–2110, Jul. 2016.

[29] R. Chikhi and P. Medvedev, "Informed and automated k-mer size selection for genome assembly," *Bioinformatics*, vol. 30, no. 1, pp. 31–37, Jan. 2014.

[30] P. A. Pevzner, H. Tang, and M. S. Waterman, "An Eulerian path approach to DNA fragment assembly," *Proc. Nat. Acad. Sci. USA*, vol. 98, no. 17, pp. 9748–9753, Aug. 2001.

[31] M. Chaisson, P. Pevzner, and H. Tang, "Fragment assembly with short reads," *Bioinformatics*, vol. 20, no. 13, pp. 2067–2074, Sep. 2004.

[32] P. A. Pevzner, "1-tuple DNA sequencing: Computer analysis," *J. Biomol. Struct. Dyn.*, vol. 7, no. 1, pp. 63–73, Aug. 1989.

[33] M. J. Chaisson and P. A. Pevzner, "Short read fragment assembly of bacterial genomes," *Genome Res.*, vol. 18, no. 2, pp. 324–330, Feb. 2008.

[34] D. R. Zerbino and E. Birney, "Velvet: Algorithms for *de novo* short read assembly using de Bruijn graphs," *Genome Res.*, vol. 18, no. 5, pp. 821–829, May 2008.

[35] M. J. Chaisson, D. Brinza, and P. A. Pevzner, "*De novo* fragment assembly with short mate-paired reads: Does the read length matter?" *Genome Res.*, vol. 19, no. 2, pp. 336–346, Feb. 2009.

[36] S. Gnerre *et al.*, "High-quality draft assemblies of mammalian genomes from massively parallel sequence data," *Proc. Nat. Acad. Sci. USA*, vol. 108, no. 4, pp. 1513–1518, Jan. 2011.

[37] R. Li *et al.*, "*De novo* assembly of human genomes with massively parallel short read sequencing," *Genome Res.*, vol. 20, no. 2, pp. 265–272, Feb. 2010.

[38] Z. Iqbal, M. Caccamo, I. Turner, P. Flicek, and G. McVean, "*De novo* assembly and genotyping of variants using colored de Bruijn graphs," *Nature Gen.*, vol. 44, no. 2, pp. 226–232, Feb. 2012.

[39] P. Medvedev, S. Pham, M. Chaisson, G. Tesler, and P. Pevzner, "Paired de Bruijn graphs: A novel approach for incorporating mate pair information into genome assemblers," *J. Comput. Biol.*, vol. 18, no. 11, pp. 1625–1634, Nov. 2011.

[40] S. K. Pham, D. Antipov, A. Sirotkin, G. Tesler, P. A. Pevzner, and M. A. Alekseyev, "Pathset graphs: A novel approach for comprehensive utilization of paired reads in genome assembly," *J. Comput. Biol.*, vol. 20, no. 4, pp. 359–371, Apr. 2013.

[41] A. Bankevich *et al.*, "SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing," *J. Comput. Biol.*, vol. 19, no. 5, pp. 455–477, May 2012.

[42] J. T. Simpson, K. Wong, S. D. Jackman, J. E. Schein, S. J. M. Jones, and I. Birol, "ABySS: A parallel assembler for short read sequence data," *Genome Res.*, vol. 19, no. 6, pp. 1117–1123, Jun. 2009.

[43] S. Boisvert, F. Laviolette, and J. Corbeil, "Ray: Simultaneous assembly of reads from a mix of high-throughput sequencing technologies," *J. Comput. Biol.*, vol. 17, no. 11, pp. 1519–1533, Nov. 2010.

[44] J. A. Chapman, I. Ho, S. Sunkara, S. Luo, G. P. Schroth, and D. S. Rokhsar, "Meraculous: *De novo* genome assembly with short paired-end reads," *PLoS ONE*, vol. 6, no. 8, p. e23501, Aug. 2011.

[45] E. Georganas *et al.*, "HipMer: An extreme-scale *de novo* genome assembler," in *Proc. Int. Conf. High Perform. Comput. Netw. Storage Anal.*, 2015, pp. 1–11.

[46] V. Deshpande, E. D. K. Fung, S. Pham, and V. Bafna, "Cerulean: A hybrid assembly using high throughput short and long reads," in *Algorithms in Bioinformatics*. Berlin, Germany: Springer-Verlag, 2013, pp. 349–363.

[47] D. Antipov, A. Korobeynikov, J. S. McLean, and P. A. Pevzner, "hybridSPAdes: An algorithm for hybrid assembly of short and long reads," *Bioinformatics*, Nov. 2015.

[48] A. Bankevich and P. A. Pevzner, "TruSPAdes: Barcode assembly of TruSeq synthetic long reads," *Nature Methods*, vol. 13, no. 3, pp. 248–250, Mar. 2016.

[49] M. Boetzer, C. V. Henkel, H. J. Jansen, D. Butler, and W. Pirovano, "Scaffolding pre-assembled contigs using SSPACE," *Bioinformatics*, vol. 27, no. 4, pp. 578–579, Feb. 2011.

[50] N. Donmez and M. Brudno, "SCARPA: Scaffolding reads with practical algorithms," *Bioinformatics*, vol. 29, no. 4, pp. 428–434, Feb. 2013.

[51] K. Sahlin, F. Vezzi, B. Nystedt, J. Lundeberg, and L. Arvestad, "BESST—Efficient scaffolding of large fragmented assemblies," *BMC Bioinformatics*, vol. 15, p. 281, Aug. 2014.

[52] S. Gao, W.-K. Sung, and N. Nagarajan, "Opera: Reconstructing optimal genomic scaffolds with high-throughput paired-end sequences," *J. Comput. Biol.*, vol. 18, no. 11, pp. 1681–1691, Nov. 2011.

[53] R. L. Warren *et al.*, "LINKS: Scalable, alignment-free scaffolding of draft genomes with long reads," *GigaScience*, vol. 4, p. 35, Aug. 2015.

[54] J. M. Shelton *et al.*, "Tools and pipelines for BioNano data: Molecule assembly pipeline and FASTA super scaffolding tool," *BMC Genomics*, vol. 16, no. 1, p. 734, Sep. 2015.

[55] A. Mortazavi *et al.*, "Scaffolding *a Caenorhabditis* nematode genome with RNA-seq," *Genome Res.*, vol. 20, no. 12, pp. 1740–1747, Dec. 2010.

[56] S. V. Zhang, L. Zhuo, and M. W. Hahn, "AGOUTI: Improving genome assembly and annotation using transcriptome data," *GigaScience*, vol. 5, no. 1, p. 31, Jul. 2016.

[57] R. Luo *et al.*, "SOAPdenovo2: An empirically improved memory-efficient short-read *de novo* assembler," *GigaScience*, vol. 1, no. 1, p. 18, Dec. 2012.

[58] D. Paulino, R. L. Warren, B. P. Vandervalk, A. Raymond, S. D. Jackman, and I. Birol, "Sealer: A scalable gap-closing application for finishing draft genomes," *BMC Bioinf.*, vol. 16, no. 1, p. 230, 2015.

[59] J. R. Miller, S. Koren, and G. Sutton, "Assembly algorithms for next-generation sequencing data," *Genomics*, vol. 95, no. 6, pp. 315–327, Jun. 2010.

[60] D. A. Earl *et al.*, "Assemblathon 1: A competitive assessment of *de novo* short read assembly methods," *Genome Res.*, vol. 21, no. 12, pp. 2224–2241, Dec. 2011.

[61] A. M. Phillippy, M. C. Schatz, and M. Pop, "Genome assembly forensics: Finding the elusive mis-assembly," *Genome Biol.*, vol. 9, no. 3, p. R55, Mar. 2008.

[62] I. M. Dew, B. Walenz, and G. Sutton, "A tool for analyzing mate pairs in assemblies (TAMPA)," *J. Comput. Biol.*, vol. 12, no. 5, pp. 497–513, Jun. 2005.

[63] M. Hunt, T. Kikuchi, M. Sanders, C. Newbold, M. Berriman, and T. D. Otto, "REAPR: A universal tool for genome assembly evaluation," *Genome Biol.*, vol. 14, no. 5, p. R47, May 2013.

[64] F. Vezzi, G. Narzisi, and B. Mishra, "Reevaluating assembly evaluations with feature response curves: GAGE and assemblathons," *PLoS ONE*, vol. 7, no. 12, p. e52210, Dec. 2012.

[65] B. Teague *et al.*, "High-resolution human genome structure by single-molecule analysis," *Proc. Nat. Acad. Sci. USA*, vol. 107, no. 24, pp. 10848–10853, Jun. 2010.

[66] M. Pendleton *et al.*, "Assembly and diploid architecture of an individual human genome via single-molecule technologies," *Nature Methods*, vol. 12, no. 8, pp. 780–786, Aug. 2015.

[67] K. R. Bradnam *et al.*, "Assemblathon 2: Evaluating *de novo* methods of genome assembly in three vertebrate species," *GigaScience*, vol. 2, no. 1, p. 10, Jul. 2013.

[68] S. L. Salzberg *et al.*, "GAGE: A critical evaluation of genome assemblies and assembly algorithms," *Genome Res.*, vol. 22, no. 3, pp. 557–567, Mar. 2012.

[69] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool," *J. Molecular Biol.*, vol. 215, no. 3, pp. 403–410, Oct. 1990.

[70] A. Gurevich, V. Saveliev, N. Vyahhi, and G. Tesler, "QUAST: Quality assessment tool for genome assemblies," *Bioinformatics*, vol. 29, no. 8, pp. 1072–1075, Apr. 2013.

[71] S. Clark, R. Egan, P. I. Frazier, and Z. Wang, "ALE: A generic assembly likelihood evaluation framework for assessing the accuracy of genome and metagenome assemblies," *Bioinformatics*, vol. 29, no. 4, pp. 435–443, Feb. 2013.

[72] M. Ghodsi *et al.*, "*De novo* likelihood-based measures for comparing genome assemblies," *BMC Res. Notes*, vol. 6, p. 334, Aug. 2013.

[73] A. Rahman and L. Pachter, "CGAL: Computing genome assembly likelihoods," *Genome Biol.*, vol. 14, no. 1, p. R8, Jan. 2013.

[74] E. W. Myers, "Toward simplifying and accurately formulating fragment assembly," *J. Comput. Biol.*, vol. 2, no. 2, pp. 275–290, Summer 1995.

[75] T. Tatusova, S. Ciufo, B. Fedorov, K. O'Neill, and I. Tolstoy, "RefSeq microbial genomes database: New representation and annotation strategy," *Nucl. Acids Res.*, vol. 42, no. D1, pp. D553–D559, Jan. 2014.

[76] K. D. Pruitt *et al.*, "RefSeq: An update on mammalian reference sequences," *Nucl. Acids Res.*, vol. 42, no. D1, pp. D756–D763, Jan. 2014.

[77] K. D. Pruitt, T. Tatusova, and J. Ostell, "The reference sequence (RefSeq) project, the NCBI handbook [Internet]," Nat. Library Med. (US), Nat. Center Biotechnol. Inf., Bethesda, MD, USA, Tech. Rep., 2002.

[78] G. Parra, K. Bradnam, and I. Korf, "CEGMA: A pipeline to accurately annotate core genes in eukaryotic genomes," *Bioinformatics*, vol. 23, no. 9, pp. 1061–1067, May 2007.

[79] L. Florea, A. Souvorov, T. S. Kalbfleisch, and S. L. Salzberg, "Genome assembly has a major impact on gene content: A comparison of annotation in two *Bos taurus* assemblies," *PLoS ONE*, vol. 6, no. 6, p. e21400, Jun. 2011.

[80] E. S. Lander *et al.*, "Initial sequencing and analysis of the human genome," *Nature*, vol. 409, no. 6822, pp. 860–921, Feb. 2001.

[81] J. C. Venter *et al.*, "The sequence of the human genome," *Science*, vol. 291, no. 5507, pp. 1304–1351, Feb. 2001.

[82] International Human Genome Sequencing Consortium, "Finishing the euchromatic sequence of the human genome," *Nature*, vol. 431, no. 7011, pp. 931–945, Oct. 2004.

[83] W. Jang *et al.*, "Linking the human cytogenetic map with nucleotide sequence: The CCAP clone set," *Cancer Genet. Cytogenet.*, vol. 168, no. 2, pp. 89–97, Jul. 2006.

[84] J. D. McPherson *et al.*, "A physical map of the human genome," *Nature*, vol. 409, no. 6822, pp. 934–941, Feb. 2001.

[85] P. Green, "Against a whole-genome shotgun," *Genome Res.*, vol. 7, no. 5, pp. 410–417, May 1997.

[86] J. L. Weber and E. W. Myers, "Human whole-genome shotgun sequencing," *Genome Res.*, vol. 7, no. 5, pp. 401–409, May 1997.

[87] Y. Xue *et al.*, "Adaptive evolution of *UGT2B17* copy-number variation," *Amer. J. Human Gen.*, vol. 83, no. 3, pp. 337–346, Sep. 2008.

[88] M. C. Zody *et al.*, "Evolutionary toggling of the *MAPT* 17q21.31 inversion region," *Nature Gen.*, vol. 40, no. 9, pp. 1076–1083, Sep. 2008.

[89] M. Y. Dennis *et al.*, "Evolution of human-specific neural *SRGAP2* genes by incomplete segmental duplication," *Cell*, vol. 149, no. 4, pp. 912–922, May 2012.

[90] F. Antonacci *et al.*, "Palindromic *GOLGA8* core duplicons promote chromosome 15q13.3 microdeletion and evolutionary instability," *Nature Gen.*, vol. 46, no. 12, pp. 1293–1302, Dec. 2014.

[91] D. M. Church *et al.*, "Modernizing reference genome assemblies," *PLoS Biol.*, vol. 9, no. 7, p. e1001091, Jul. 2011.

[92] D. M. Church *et al.*, "Extending reference assembly models," *Genome Biol.*, vol. 16, p. 13, Jan. 2015.

[93] V. A. Schneider *et al.*, "Evaluation of GRCh38 and *de novo* haploid genome assemblies demonstrates the enduring quality of the reference assembly," *bioRxiv*, p. 072116, Jan. 2016.

[94] J. M. Kidd *et al.*, "Mapping and sequencing of structural variation from eight human genomes," *Nature*, vol. 453, pp. 56–64, May 2008.

[95] K. M. Steinberg *et al.*, "Single haplotype assembly of the human genome from a hydatidiform mole," *Genome Res.*, vol. 24, no. 12, pp. 2066–2076, Dec. 2014.

[96] C. T. Watson *et al.*, "Complete haplotype sequence of the human immunoglobulin heavy-chain variable, diversity, and joining genes and characterization of allelic and copy-number variation," *Amer. J. Human Gen.*, vol. 92, no. 4, pp. 530–546, Apr. 2013.

[97] P. A. Kitts *et al.*, "Assembly: A resource for assembled genomes at NCBI," *Nucl. Acids Res.*, Nov. 2015.

[98] C. Ye, C. Hill, S. Wu, J. Ruan, and Z. Ma, "DBG2OLC: Efficient assembly of large genomes using long erroneous reads of the third generation sequencing technologies," Oct. 10, 2014 [Online]. Available: https://arxiv.org/abs/1410.2801

[99] J. F. Denton, J. Lugo-Martinez, A. E. Tucker, D. R. Schrider, W. C. Warren, and M. W. Hahn, "Extensive error in the number of genes inferred from draft genome assemblies," *PLoS Comput. Biol.*, vol. 10, no. 12, p. e1003998, 2014.

[100] F. Vezzi, G. Narzisi, and B. Mishra, "Feature-by-feature—Evaluating *de novo* sequence assembly," *PLoS ONE*, vol. 7, no. 2, p. e31002, 2012.

[101] D. M. Church *et al.*, "Extending reference assembly models," *Genome Biol.*, vol. 16, p. 13, Jan. 2015.

[102] B. Paten, A. Novak, and D. Haussler, "Mapping to a reference genome structure," Apr. 20, 2014 [Online]. Available: https://arxiv.org/abs/1404.5010

[103] S. Marcus, H. Lee, and M. C. Schatz, "SplitMEM: A graphical algorithm for pan-genome analysis with suffix skips," *Bioinformatics*, vol. 30, no. 24, pp. 3476–3483, Dec. 2014.

[104] E. Lieberman-Aiden *et al.*, "Comprehensive mapping of long-range interactions reveals folding principles of the human genome," *Science*, vol. 326, no. 5950, pp. 289–293, Oct. 2009.

[105] R. Kalhor, H. Tjong, N. Jayathilaka, F. Alber, and L. Chen, "Genome architectures revealed by tethered chromosome conformation capture and population-based modeling," *Nature Biotechnol.*, vol. 30, no. 1, pp. 90–98, Jan. 2012.

[106] J. R. Dixon *et al.*, "Topological domains in mammalian genomes identified by analysis of chromatin interactions," *Nature*, vol. 485, no. 7398, pp. 376–380, May 2012.

[107] J. N. Burton, A. Adey, R. P. Patwardhan, R. Qiu, J. O. Kitzman, and J. Shendure, "Chromosome-scale scaffolding of *de novo* genome assemblies based on chromatin interactions," *Nature Biotechnol.*, vol. 31, no. 12, pp. 1119–1125, Dec. 2013.

[108] N. H. Putnam *et al.*, "Chromosome-scale shotgun assembly using an *in vitro* method for long-range linkage," Feb. 18, 2015 [Online]. Available: https://arxiv.org/abs/1502.05331

[109] M. Akeson, D. Branton, G. Church, and D. W. Deamer, "Characterization of individual polymer molecules based on monomer-interface interactions," U.S. Patent 20120160687 A1, Jun. 28, 2012.

[110] J. Quick *et al.*, "Rapid draft sequencing and real-time nanopore sequencing in a hospital outbreak of *Salmonella*," *Genome Biol.*, vol. 16, p. 114, May 2015.

ABOUT THE AUTHORS

Karyn Meltz Steinberg received the B.S. degree in anthropology, human biology, 2001) from Northwestern University, Evanston, IL, USA, in 2001 and the Ph.D. degree in biological and biomedical sciences from Emory University, Atlanta, GA, USA, in 2009.

She completed a postdoctoral fellowship at the University of Washington, Seattle, WA, USA, before joining the faculty at McDonnell Genome Institute, Washington University School of Medicine, St. Louis, MO, USA.

Dr. Meltz Steinberg has received numerous awards including a National Science Foundation Graduate Research Fellowship and a Ruth L. Kirschstein National Research Service Award Postdoctoral Fellowship, and she was a semi-finalist for the Charles J. Epstein Post-Doctoral Trainee Award for Excellence in Human Genetics Research through the American Society of Human Genetics.

Valerie A. Schneider received the B.S. degree in biology from Cornell University, Ithaca, NY, USA, in 1994 and the PhD degree in biological and biomedical sciences from Harvard University, Cambridge, MA, USA, in 2001.

From 2001 to 2007, she was a Postdoctoral Fellow at the University of Pennsylvania, Philadelphia, PA, USA. Since 2008, she has been a Staff Scientist at the National Center for Biotechnology Information (NCBI), Bethesda, MD, USA. Her research interests include genomic assemblies, sequence variation, and the development of visualization tools for genomic analyses.

Dr. Schneider is a member of the International Committee on Standardized Genetic Nomenclature for Mice. She has received several honors, including being awarded a Ruth L. Kirschstein National Research Service Award Postdoctoral Fellowship and the Holtzer Prize for outstanding postdoctoral research, from the University of Pennsylvania.

Can Alkan received the B.S. degree in computer engineering from Bilkent University, Ankara, Turkey, in 2000 and the Ph.D. degree in computer science from Case Western University, Cleveland, OH, USA, in 2005.

He completed a postdoctoral fellowship at the University of Washington, Seattle, WA, USA, before joining the faculty at the Department of Computer Engineering, Bilkent University, Ankara, Turkey.

Dr. Alkan has received numerous awards including the Young Investigator Award from the Science Academy of Turkey and the Incentive Award from the Scientific and Technological Research Council of Turkey.

Michael J. Montague received the B.S. degree in biology from Boston College, Chestnut Hill, MA, USA, in 2000 and the Ph.D. degree in biological anthropology from New York University, New York, NY, USA, in 2011.

He completed a postdoctoral fellowship at the McDonnell Genome Institute, School of Medicine, Washington University at St. Louis, St. Louis, MO, USA and currently works as a Postdoctoral Researcher in the Department of Neuroscience, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA.

Dr. Montague has received numerous awards including a National Science Foundation Graduate Research Fellowship, a Wenner-Gren Dissertation Fieldwork Grant, and a Sokal Research Award for Predoctoral Students in the Sciences.

Wesley C. Warren received the B.S. degree in animal science from Oklahoma State University, Stillwater, OK, USA, in 1984, the M.S. degree in reproductive physiology from Clemson University, Clemson, SC, USA, in 1986, and the Ph.D. degree in molecular endocrinology from University of Missouri, Columbia, MO, USA, in 1990.

He was a Postdoctoral Fellow at G.D. Searle, Skokie, IL, USA, from 1990 to 1992 and a member of the Molecular Biology and Genomics groups at Monsanto, St. Louis, MO, USA, from 1992 to 2000. From 2000 to 2001, he was Senior Director of Operations at Incyte Genomics. Since 2002, he has been an Assistant Director of McDonnell Genome Institute and an Associate Professor of Genetics at Washington University in St. Louis, St. Louis, MO, USA. His research interests include *de novo* genome assemblies and comparative genomics.

Deanna M. Church received the B.A. degree in liberal arts from the University of Virginia, Charlottesville, VA, USA, in 1990 and the Ph.D. degree in biological sciences from the University of California Irvine, Irvine, CA, USA, with Dr. J. Wasmuth.

She did postdoctoral training in developmental biology with Dr. J. Rossant at the Samuel Lunenfeld Research Institute, Mount Sinai Hospital, Toronto, ON, Canada. Currently, she is the Senior Director of Applications at 10x Genomics, Pleasanton, CA, USA. In this role, she leads a diverse group of scientists who are developing approaches for improved genome analysis, using linked-reads, as well as expanding the application space of single-cell transcriptome profiling. Previously, she was Senior Director of Genomics and Content at Personalis, Inc., Menlo Park, CA, USA, where she helped advance the field of genomics-based clinical diagnostics. Prior to that, she was a staff scientist at the National Center for Biotechnology Information (NCBI), Bethesda, MD, USA, where she oversaw several projects concerning managing and displaying genomic data, including dbVar, a database of structural variation, the NCBI Variation Viewer, the NCBI Map Viewer, the Clone database, and the NCBI Remap service. She was also a founding member of the Genome Reference Consortium (GRC), an international group charged with improving the reference assembly for humans and other model organisms and was an author on the two seminal manuscripts describing the human and mouse genome sequences. She has experience in molecular biology, genetics, genomics, and bioinformatics.

Richard K. Wilson received the A.B. degree in microbiology from Miami University, Miami, FL, USA, in 1981 and the Ph.D. degree in biochemistry from the University of Oklahoma, Norman, OK, USA, in 1986.

He was a Research Fellow at the California Institute of Technology, Pasadena, CA, USA (1986–1990) before joining the faculty of the Department of Genetics, Washington University School of Medicine, St. Louis, MO, USA, in 1990–2016. He is currently a Professor of Pediatrics at The Ohio State University College of Medicine, Columbus, OH, USA and the inaugural Executive Director of the Institute for Genomic Medicine at Nationwide Children's Hospital, Columbus, OH, USA. He is an internationally recognized expert in molecular genetics and large-scale genome analysis. He was a member of the team that developed automated DNA sequencing, and subsequently led the team at Washington University in St. Louis that was the first to sequence the genome of a multicellular animal (the roundworm *C. elegans*). After playing a significant role in the Human Genome Project, his laboratory was the first to sequence the genome of a cancer patient and identify the somatic mutations responsible for the patient's disease. His group's subsequent studies have revealed numerous key characteristics of several adult and pediatric cancer types related to relapse risk, metastatic disease, and mechanisms of acquired resistance.