New advances in sequence assembly

Adam M. Phillippy

National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland 20892, USA

It may be hard to believe, but the idea of sequence assembly is around 40 years old. Consider this pair of quotes from Rodger Staden (Staden 1979):

> "With modern fast sequencing techniques and suitable computer programs it is now possible to sequence whole genomes without the need of restriction maps."
>
> "If the 5′ end of the sequence from one gel reading is the same as the 3′ end of the sequence from another the data is said to overlap. If the overlap is of sufficient length to distinguish it from being a repeat in the sequence the two sequences must be contiguous. The data from the two gel readings can then be joined to form one longer continuous sequence."

Replace "gel reading" with "read" and these sentences would go unnoticed in the introduction of any paper today. Here you can also see the birth of jargon that now pervades the field: *overlaps* between *reads* form *contigs* (contiguous sequences). Just a few months later, Gingeras et al. (1979) described "Computer programs for the assembly of DNA sequences." It all sounds so modern, until the discussion mentions FORTRAN code stored on magnetic tapes.

How, then, can we fill an entire special issue of *Genome Research* with "new advances" so many years later? To me, this reflects the beauty of the problem—simple enough to be stated in a single paragraph, yet complex enough to sustain a field of research for decades. This dichotomy is common to many famous computational problems; indeed, mathematical formulations of sequence assembly fall into a class of problems known as "NP-hard" that do not admit an easy solution (Medvedev et al. 2007).

There is another reason for continued advances in sequence assembly—advances in sequencing technology. As evident from the Staden quotes above, the first assembly methods were motivated by the invention of DNA sequencing and gel electrophoresis "readings" (Sanger and Coulson 1975; Maxam and Gilbert 1977). These early sequencing and assembly methods were applied to viruses with genomes of only a few kilobases (Sanger et al. 1978). As the sequencing technology was later commercialized and scaled, it became possible to assemble the 1-Mb genome of a free-living bacterium (Fleischmann et al. 1995), the 120-Mb genome of the fruit fly (Adams et al. 2000), and ultimately the 3-Gb human genome (Venter et al. 2001). These advances in scale required parallel computational advances, embodied by the tools that assembled these early genomes—TIGR Assembler (Sutton et al. 1995) and Celera Assembler (Myers et al. 2000). A similar theme continued through the 2000s, as new sequencing technology such as 454 (Margulies et al. 2005) and Illumina (formerly Solexa) (Bentley et al. 2008) required rethinking the assembly problem. The abrupt transition away from Sanger reads to the much shorter Illumina reads shifted the field toward de Bruijn graph assemblers (Pevzner et al. 2001) such as Velvet (Zerbino and Birney 2008) and ABySS (Simpson et al. 2009). Most recently, with the advent of longer, single-molecule sequencing reads from Pacific Biosciences (PacBio) (Eid et al. 2009), the field returned to overlap graphs (Myers 1995) and adaptations of Celera Assembler (Chin et al. 2013; Koren et al. 2013).

Presently, the number of available technologies has only grown. The papers in this issue include sequencing data from four platforms—Illumina, PacBio, Oxford Nanopore, and Sanger—and multiple technologies for constructing long-range scaffolds: paired reads, linked reads, optical mapping, proximity ligation, and physical mapping. This glut of technologies has spurred interest in determining the most effective approach to reconstruct whole genomes.

The low cost of short-read sequencing compared to Sanger has driven a wide expansion in the number of genomes sequenced, but with a sharp reduction in contig and scaffold lengths. An emerging trend is to combine cost-effective Illumina sequencing with clever library preparation techniques designed to improve assembly continuity. One powerful example is chromatin conformation capture via proximity ligation and high-throughput sequencing (Hi-C) (Lieberman-Aiden et al. 2009). This family of methods generates a familiar paired-read data type (two reads separated by some distance) but from a distribution of sizes that can span megabases. This data can be used to group contigs by chromosome, reconstruct chromosome-length scaffolds, and phase haplotypes (Burton et al. 2013; Kaplan and Dekker 2013; Selvaraj et al. 2013). In this issue, Rice et al. (2017) demonstrates a related approach, using in vitro reconstituted chromatin and Illumina sequencing to assemble the American alligator genome. Another approach to boosting short reads uses high-throughput barcoding to tag groups of "linked reads" that all originate from a larger, single molecule of DNA. For this new data type, Weisenfeld et al. (2017) introduces a new assembler, Supernova, for the de novo assembly of diploid human genomes from linked reads. Additionally, Jackman et al. (2017) describes a new version of the ABySS assembler and explores linked reads and optical mapping for improved scaffolding.

Although these short-read library preparation methods can extend scaffolds to span entire chromosomes, they lack the finer resolution required to improve contig lengths. Instead, the biggest gains in contig lengths have come from single-molecule sequencing. First from PacBio and most recently from Oxford Nanopore, these technologies can generate reads exceeding 10 kb, orders of magnitude longer than Illumina. Critically, 10-kb reads are longer than the most common repeats in both microbial and vertebrate genomes and can therefore generate highly continuous assemblies. In fact, the complete reconstruction of bacterial genomes—a process that used to require teams of people—is now automated and routine. However, the massive read lengths and increased error rate of these new technologies have also required updated assembly methods. This issue includes three new assembly tools

Corresponding author: adam.phillippy@nih.gov

Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.223057.117. Freely available online through the *Genome Research* Open Access option.

designed specifically for long-read PacBio and Nanopore data: Canu (Koren et al. 2017), HINGE (Kamath et al. 2017), and Racon (Vaser et al. 2017).

Combining single-molecule sequencing with complementary technologies has also become a common strategy. Fan et al. (2017) demonstrates improved accuracy for human structural variant calling using a combination of PacBio and Illumina. For de novo assembly, two new studies look at plant genome assembly, with Zimin et al. (2017) combining PacBio and Illumina data to assemble the highly repetitive grass *Aegilops tauschii*, and Jiao et al. (2017) combining PacBio with proximity ligation and optical mapping to assemble relatives of the model species, *Arabadopsis thaliana*. In the latter case, the combination of PacBio reads and long-range scaffolding techniques enabled multi-megabase contigs and scaffolds spanning entire chromosome arms.

So, what is gained from improved de novo assemblies? High-quality assemblies can reveal repeat structures and structural variation otherwise missed by short-read resequencing. For de novo projects, short-read assemblies are fragmented, because the overlaps between reads are not long enough to distinguish between common repeats. These repeat sequences are left unassembled, breaking contigs and hindering analysis. With the emergence of long sequencing reads has come a renewed interest in repetitive sequences, which can be properly analyzed for the first time. This includes detailed analysis of highly repetitive satellite sequence in flies (Khosta et al. 2017) and birds (Weissensteiner et al. 2017), paving the way for functional studies in areas of the genome not previously accessible. Long-read assembly is even revealing new variation in the human genome, and Huddleston et al. (2017) highlights the importance of long-read sequencing and haplotype resolution for accurate structural variant detection.

Ultimately, the goal of genome assembly is a gapless, haplotype-resolved reconstruction, but these genomic jigsaw puzzles are so difficult that we have not yet finished the human genome. This issue marks the 38th build of the human reference sequence (Schneider et al. 2017), which still contains more than 800 gaps after decades of work and billions of dollars spent. But there is hope on the horizon. Progress over the past five years has been swift, driven by new technology. De novo assemblies of humans (Seo et al. 2016) and other vertebrates (Bickhart et al. 2017) are approaching reference quality by combining technologies such as PacBio, Illumina, optical mapping, linked reads, and proximity ligation; and new phasing methods can now recover chromosome-scale haplotype blocks from this data (Edge et al. 2017). With these latest techniques, only the largest segmental duplications and heterochromatic regions remain a challenge.

Future advances in technology may overcome these remaining hurdles. Reflecting on the earlier Staden quote, "If the overlap is of sufficient length to distinguish it from being a repeat …," as soon as a sequencing technology can produce good enough reads, such that all regions of the genome can be uniquely distinguished, genome assembly will become trivial. Although the minimum combination of read accuracy and length required to complete the human genome is currently unknown, I suspect we are getting close. Most recently, Nanopore sequencing reads approaching 1 Mbp were reported (http://lab.loman.net/2017/03/09/ultrareads-for-nanopore/), and it is imaginable that further technology advances will enable the complete assembly of a diploid human within a few years.

Luckily, even if sequence assembly is made obsolete by new technology, there will be plenty of work left for the bioinformaticians. Low-cost, complete genomes will enable new and powerful comparative genomics studies, requiring scalable methods for analyzing whole genomes. Many of these methods share much in common with the de Bruijn and overlap graphs of genome assembly. Illustrating these connections, Paten et al. (2017) provides a state-of-the-art overview on the topic of genome graphs and their application to read mapping, variant calling, and haplotype determination. These graph structures allow reference genomes to evolve beyond a single, linear representation to capture the full diversity of a population. Hopefully, as continued advances in technology allow us to spend less time assembling genomes, we can spend more time exploring their wonderful evolution and functional complexity.

Competing interest statement

A.M.P. served as a guest editor for this issue of *Genome Research*, had access to all papers prior to publication, and coauthored two included papers.

References

Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, Scherer SE, Li PW, Hoskins RA, Galle RF, et al. 2000. The genome sequence of *Drosophila melanogaster*. *Science* **287:** 2185–2195.

Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR, et al. 2008. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456:** 53–59.

Bickhart DM, Rosen BD, Koren S, Sayre BL, Hastie AR, Chan S, Lee J, Lam ET, Liachko I, Sullivan ST, et al. 2017. Single-molecule sequencing and chromatin conformation capture enable de novo reference assembly of the domestic goat genome. *Nat Genet* doi: 10.1038/ng.3802.

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:** 1119–1125.

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

Edge P, Bafna V, Bansal V. 2017. HapCUT2: robust and accurate haplotype assembly for diverse sequencing technologies. *Genome Res* (this issue). doi: 10.1101/gr.213462.116.

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

Fan X, Chaisson M, Nakhleh L, Chen K. 2017. HySA: a Hybrid Structural variant Assembly approach using next-generation and single-molecule sequencing technologies. *Genome Res* (this issue). doi: 10.1101/gr.214767.116.

Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA, Merrick JM, et al. 1995. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269:** 496–512.

Gingeras TR, Milazzo JP, Sciaky D, Roberts RJ. 1979. Computer programs for the assembly of DNA sequences. *Nucleic Acids Res* **7:** 529–545.

Huddleston J, Chaisson MJ, Meltz Steinberg K, Warren W, Hoekzema K, Gordon D, Graves-Lindsay TA, Munson KM, Kronenberg ZN, Vives L, et al. 2017. Discovery and genotyping of structural variation from long-read haploid genome sequence data. *Genome Res* (this issue). doi: 10.1101/gr.214007.116.

Jackman SD, Vandervalk BP, Mohamadi H, Chu J, Yeo S, Hammond SA, Jahesh G, Khan H, Coombe L, Warren RL, et al. 2017. ABySS 2.0: resource-efficient assembly of large genomes using a Bloom filter. *Genome Res* (this issue). doi: 10.1101/gr.214346.116.

Jiao WB, Garcia Accinelli G, Hartwig B, Kiefer C, Baker D, Severing E, Willing EM, Piednoel M, Woetzel S, Madrid-Herrero E, et al. 2017. Improving and correcting the contiguity of long-read genome assemblies of three plant species using optical mapping and chromosome conformation capture data. *Genome Res* (this issue). doi: 10.1101/gr.213652.116.

Kamath GM, Shomorony I, Xia F, Courtade TA, Tse DN. 2017. HINGE: long-read assembly achieves optimal repeat resolution. *Genome Res* (this issue). doi: 10.1101/gr.216465.116.

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

Khost DE, Eickbush DG, Larracuente AM. 2017. Single-molecule sequencing resolves the detailed structure of complex satellite DNA loci in *Drosophila melanogaster*. *Genome Res* (this issue). doi: 10.1101/gr.213512.116.

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

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* (this issue). doi: 10.1101/gr.215087.116.

Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, et al. 2009. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326:** 289–293.

Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z, et al. 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437:** 376–380.

Maxam AM, Gilbert W. 1977. A new method for sequencing DNA. *Proc Natl Acad Sci* **74:** 560–564.

Medvedev P, Georgiou K, Myers G, Brudno M. 2007. Computability of models for sequence assembly. In *Algorithms in bioinformatics*. *WABI 2007* (ed. Giancarlo R, Hannenhalli S), Vol. 4645 of *Lecture Notes in Computer Science*. Springer, Berlin, Heidelberg.

Myers EW. 1995. Toward simplifying and accurately formulating fragment assembly. *J Comput Biol* **2:** 275–290.

Myers EW, Sutton GG, Delcher AL, Dew IM, Fasulo DP, Flanigan MJ, Kravitz SA, Mobarry CM, Reinert KH, Remington KA, et al. 2000. A whole-genome assembly of *Drosophila*. *Science* **287:** 2196–2204.

Paten B, Novak AM, Eizenga JM, Garrison E. 2017. Genome graphs and the evolution of genome inference. *Genome Res* (this issue). doi: 10.1101/gr.214155.116.

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

Rice ES, Kohno S, St. John J, Pham S, Howard J, Lareau LF, O'Connell BL, Hickey G, Armstrong J, Deran A, et al. 2017. Improved genome assembly of American alligator genome reveals conserved architecture of estrogen signaling. *Genome Res* (this issue). doi: 10.1101/gr.213595.116.

Sanger F, Coulson AR. 1975. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J Mol Biol* **94:** 441–448.

Sanger F, Coulson AR, Friedmann T, Air GM, Barrell BG, Brown NL, Fiddes JC, Hutchison CA III, Slocombe PM, Smith M. 1978. The nucleotide sequence of bacteriophage φX174. *J Mol Biol* **125:** 225–246.

Schneider VA, Graves-Lindsay T, Howe K, Bouk N, Chen H-C, Kitts PA, Murphy TD, Pruitt KD, Thibaud-Nissen F, Albracht D, et al. 2017. Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Res* (this issue). doi: 10.1101/gr.213611.116.

Selvaraj S, R Dixon J, Bansal V, Ren B. 2013. Whole-genome haplotype reconstruction using proximity-ligation and shotgun sequencing. *Nat Biotechnol* **31:** 1111–1118.

Seo JS, Rhie A, Kim J, Lee S, Sohn MH, Kim CU, Hastie A, Cao H, Yun JY, Kim J, et al. 2016. *De novo* assembly and phasing of a Korean human genome. *Nature* **538:** 243–247.

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

Staden R. 1979. A strategy of DNA sequencing employing computer programs. *Nucleic Acids Res* **6:** 2601–2610.

Sutton GG, White O, Adams MD, Kerlavage AR. 1995. TIGR assembler: a new tool for assembling large shotgun sequencing projects. *Genome Sci Technol* **1:** 9–19.

Vaser R, Sović I, Nagarajan N, Šikić M. 2017. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res* (this issue). doi: 10.1101/gr.214270.116.

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, et al. 2001. The sequence of the human genome. *Science* **291:** 1304–1351.

Weisenfeld NI, Kumar V, Shah P, Church DM, Jaffe DB. 2017. Direct determination of diploid genome sequences. *Genome Res* (this issue). doi: 10.1101/gr.214874.116.

Weissensteiner MH, Pang AWC, Bunikis I, Höijer I, Vinnere-Petterson O, Suh A, Wolf JBW. 2017. Combination of short-read, long-read, and optical mapping assemblies reveals large-scale tandem repeat arrays with population genetic implications. *Genome Res* (this issue). doi: 10.1101/gr.215095.116.

Zerbino DR, Birney E. 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res* **18:** 821–829.

Zimin AV, Puiu D, Luo MC, Zhu T, Koren S, Marçais G, Yorke JA, Dvořák J, Salzberg SL. 2017. Hybrid assembly of the large and highly repetitive genome of *Aegilops tauschii*, a progenitor of bread wheat, with the MaSuRCA mega-reads algorithm. *Genome Res* (this issue). doi: 10.1101/gr.213405.116.