Next-Generation Sequencing Technologies

W. Richard McCombie,[1] John D. McPherson,[2] and Elaine R. Mardis[3]

[1]Genome Center, Cold Spring Harbor Laboratory, Woodbury, New York 11797

[2]Department of Biochemistry and Molecular Medicine, University of California Davis Comprehensive Cancer Center, Sacramento, California 95817

[3]Institute for Genomic Medicine at Nationwide Children's Hospital, The Ohio State University College of Medicine, The Institute for Genomic Medicine, Columbus, Ohio 43205

Correspondence: jdmcpherson@ucdavis.edu

Although DNA and RNA sequencing has a history spanning five decades, large-scale massively parallel sequencing, or next-generation sequencing (NGS), has only been commercially available for about 10 years. Nonetheless, the meteoric increase in sequencing throughput with NGS has dramatically changed our understanding of our genome and ourselves. Sequencing the first human genome as a haploid reference took nearly 10 years but now a full diploid human genome sequence can be accomplished in just a few days. NGS has also reduced the cost of generating sequence data and a plethora of sequence-based methods for probing a genome have emerged using NGS as the readout and have been applied to many species. NGS methods have also entered the medical realm and will see an increasing use in diagnosis and treatment. NGS has largely been driven by short-read generation (150 bp) but new platforms have emerged and are now capable of generating long multikilobase reads. These latter platforms enable reference-independent genome assemblies and long-range haplotype generation. Rapid DNA and RNA sequencing is now mainstream and will continue to have an increasing impact on biology and medicine.

HISTORY OF DNA SEQUENCING

Sanger Enzymatic Sequencing

Although the Sanger dideoxynucleotide sequencing method was introduced in 1977 (Sanger et al. 1977), other enzymatic sequencing methods were devised and published in the same time frame, including partial ribosubstitution sequencing of Barnes (1978), the plus and minus method of Sanger (Sanger and Coulson 1975) and the chemical cleavage method (Maxam and Gilbert 1977). Dideoxynucleotide sequencing emerged from the interest in Sanger's laboratory to devise an enzymatic approach to DNA sequencing, and the studies of *Escherichia coli* DNA polymerase I and its interactions with various nucleotides and nucleotide analogs as substrates or inhibitors in Kornberg's laboratory. At its essence, the enzymatic DNA sequencing reaction mimics many aspects of DNA replication in the cell. This is exemplified by the components of the Sanger reaction, combining the DNA polymerase with its template, primed with a synthetic oligonucleotide primer to provide a free 3′ OH for the polymerase-catalyzed addition of native nucleotides and dideoxynu-

Cite this article as *Cold Spring Harb Perspect Med* doi: 10.1101/cshperspect.a036798

cleotide analogs, the latter causing termination of the elongating nucleotide chain by preventing the addition of any further nucleotides once incorporated. By providing a carefully adjusted ratio of nucleotides (originally, one of which was radiolabeled) and specific dideoxynucleotides in each of four enzymatic primer extension reactions (one with each of the four dideoxynucleotides present), Sanger sequencing produces a pool of molecules in each reaction mix that includes some molecules that are terminated at each residue within the growing chain in which the dideoxynucleotide in that specific reaction is incorporated. When subjected to a denaturing polyacrylamide gel in a subsequent step, this produces a ladder of fragments across four lanes, each differing by one nucleotide in length. This ladder is detected when exposed to X-ray film to reveal the ladder of fragments, and read from bottom to top to derive the nucleotide sequence of each primed template (shortest to longest fragment). In summary, Sanger reactions are characterized by two discrete steps, the first is the enzymatic production of template-directed fragment ladders, each terminated by a specific dideoxynucleotide as specified by the template, and the second is an electrophoretic separation and sequence detection step.

Over time, radiolabeled (^{32}P or ^{35}S) nucleotides and film-based detection were supplanted by Hood and colleagues with the incorporation of four different fluorescent dyes enabling detection during electrophoretic separation by laser-induced fluorescent emission of each fragment on a dedicated sequencing instrument (Smith et al. 1986). Because this allowed each of the four reactions to be run in a single lane of a gel (owing to the four different emission spectra of the dyes), it eliminated most gel artifacts and allowed automated reading of the sequence ladders. Similarly, improvements in detection chemistry were accompanied by advances such as incorporating the fluorescent dye on the dideoxynucleotides rather than primers to increase flexibility, the use of polymerase enzymes with favorable characteristics that gave a more even incorporation of dideoxynucleotides and were stable at higher temperatures to alleviate secondary structure effects in high G + C content templates, and myriad other small changes that were achieved during the human genome project to make the first human reference genome possible.

Sanger Sequencing Data Analysis

Early Sanger sequencing projects focused on sequence determination for single genes or very small genomes (~5000 bases at most). The development of a computational package by Rodger Staden, developed at the Medical Research Council (MRC) in concert with Sanger's laboratory allowed the size of DNA regions or genomes sequenced to scale significantly (Staden 1979, 1984; Staden et al. 2000). The Staden Package allowed randomly sheared, small fragments of DNA from a larger original DNA source to be sequenced randomly and computationally overlapped to reform the whole sequence of the original, larger input source. The Staden Package was compiled for use on early Unix operating systems and was widely used. It provided a sequence assembly capability for input data and a viewer that permitted visual evaluation of the overlaps between fragments. Six-frame translation of assembled sequences was also output, after which the potential for open reading frames could be determined. The combination of Sanger sequencing and the Staden Package for sequence assembly created the ability to sequence genomes roughly 10 times as large as originally possible, such as phage lambda (Sanger et al. 1982) and enabled the era of genomics.

As sequencing projects became focused on longer DNA inserts (e.g., cosmid subclones) and on larger genomes, the Staden Package was supplanted by the phred/phrap/consed suite from Phil Green's laboratory (Ewing and Green 1998; Ewing et al. 1998; Gordon et al. 1998). Here, *phred* provided basecalling accuracy statistics for Sanger reads, *phrap* was a read assembly program, and *consed* was an assembly viewer and editing program. It had the ability to work with fluorescent sequence data as well. This suite provided needed scalability and was indeed the primary toolkit for sequencing and finishing of all large genomes sequenced by mapped clone-

based methods, including the international human genome project (Lander et al. 2001). In a time span of ~25 years, from the introduction of Sanger sequencing until the completion of the human genome reference sequence (1977–2004), the scalability and widespread use of Sanger sequencing had experienced significant technological advances that permitted large-scale projects to be completed. However, the effort to sequence, assemble, and annotate genomes with Sanger approaches was still a significant and expensive undertaking that required specialized equipment, expertise, and infrastructure. This scenario began to change shortly thereafter, with the introduction of the first massively parallel DNA sequencing technology in 2005 (Margulies et al. 2005).

NEXT-GENERATION SEQUENCING (NGS) TECHNOLOGIES

Fundamental Aspects of NGS

Most so-called "massively parallel" or "next-generation" sequencing methods and instruments to date have close intellectual connections to Sanger sequencing from the standpoint of their fundamental enzymological underpinnings, as will be reviewed here. The primary differentiation is that, unlike Sanger sequencing, massively parallel sequencing approaches do not decouple enzymatic nucleotide incorporation from sequence ladder separation and data acquisition. Rather, NGS instruments perform both the enzymology and data acquisition in an orchestrated and stepwise fashion, enabling sequence data to be generated from tens of thousands to billions of templates simultaneously. Hence, the term "massively parallel" refers to this enhanced data-generation capacity that has resulted in significant changes to DNA sequencing and its applications since its introduction.

Sequencing by Synthesis

Most commercial platforms using massively parallel sequencing are based on the concept of sequencing by synthesis (SBS). In essence, these methods permit nucleotide incorporation using a variety of enzymes and detection schemes that permit the corresponding instrument platform to collect data in lockstep with enzymatic synthesis on a template. Other shared concepts for these platforms include (1) fragmentation by physical shearing before sequencing of the genome or other DNA/RNA sample to be sequenced, (2) the generation of a sequencing "library" that results from the attachment of universal, platform-specific adaptors (synthesized oligonucleotides of known sequence) at each end of the template fragments to be sequenced, and (3) on-surface template amplification by virtue of library fragment hybridization to covalently attached oligonucleotides with sequence complementarity to the synthetic adaptors. Following fragment amplification, the templates are primed by virtue of unique sequences present in the synthetic adaptor, providing a free 3′ OH for enzymatic extension on the templates coupled with on-instrument data detection. The on-surface amplification establishes a fixed X–Y coordinate for each template that, in turn, permits data from nucleotide incorporation steps to be assigned to a specific template. Because SBS methods are detecting nucleotide incorporation from a population of amplified template molecules at each X–Y coordinate, and because of various types of background noise that contribute in a cumulative manner at each step in the incorporation reactions, SBS is ultimately limited in its length of sequence read ("read length") because of the increasing noise over sequential incorporation and imaging cycles. Modifications to the enzymology and nucleotide chemistry or synthesis, as well as more sensitive detectors, have yielded improved signal-to-noise over time, permitting increased read lengths. Still, SBS read lengths remain shorter than Sanger read lengths. This has impacted how the sequence data are analyzed, as will be described in the section on analytical aspects of SBS data.

SBS Detection Schema

One key difference in SBS-based massively parallel sequencing platforms is the mechanism by which nucleotide incorporation is detected. There are essentially two themes that have been

used in detection: direct fluorescence detection or indirect sensing by nucleotide incorporation reaction products. The first theme copies original Sanger sequencing most closely, by labeling the nucleotides, which are modified to only allow one base extension with fluorescent moieties that distinguish each nucleotide base by virtue of a specific fluorescent emission wavelength (e.g., Illumina). After each cycle of nucleotide addition, the dye and the extension blocking moiety are cleaved off the growing chain and the cycle is repeated. The cyclical fluorescent additions at each X–Y coordinate represent the sequence of added bases.

The second theme uses a combination of a single type of unlabeled nucleotide per incorporation cycle and a postincorporation detection step to identify which of the fragments incorporated one (or more) nucleotides. The postreaction detection is fueled by a chemical byproduct of the nucleotide incorporation reaction. As such, when multiple nucleotides of one type are adjacent in the template (i.e., sequential G nucleotides), the amount of chemical byproduct will scale accordingly, compared with the amount for a single-nucleotide incorporation. This variability in detected byproduct is, however, not infinitely linear because of the limits on the dynamic range of the detector. Hence, all instruments using this detection scheme tend to lower accuracy in mononucleotide repeat sequences. Two types of postincorporation detection schema include: (1) detection of pyrophosphate released during nucleotide incorporation(s) that reacts with firefly luciferase to produce an amount of emitted light corresponding to the number of incorporated nucleotides (e.g., Roche 454), as detected by a highly sensitive charge-coupled device (CCD) camera, or (2) detection of the pH change resulting from released hydrogen ions following nucleotide incorporation(s) as detected by an X–Y-specific miniaturized pH meter (e.g., Ion Torrent).

In general, regardless of the type of detection, each platform will process the X–Y coordinate–specific data following the sequencing instrument run completion, essentially turning the acquired signals into "reads" (i.e., the nucleotide sequence of the fragment at each coordinate). These analytical pipelines are provided by the instrument manufacturer, and perform postanalytical quality evaluation of the data, culling low-quality reads from the final output and generating metrics that permit user evaluation of the overall read data quality. Finally, the read data files are written in a specific format suitable for their use in downstream analytical pipelines.

Analytical Aspects of SBS Data

Following sequence data generation on massively parallel platforms, a wide variety of analytical steps may be pursued to provide information from the data generated. In Sanger sequencing, reads were typically derived from subcloned fragments originating from a large-insert DNA clone (cosmid, fosmid, or BAC), permitting read assembly based on sequence similarity to recapitulate the initial cloned fragment in one or more "contigs" (subassemblies of the reads). In NGS approaches, the shorter read lengths often are not suitable for read assembly, especially if the NGS library was generated from a whole genome, not a cloned fragment of that genome. This is particularly true when the genome is larger than 1–2 Mb (million bases) and is complex in nature (i.e., contains repetitive sequences). In general, short-read data are more readily interpreted by their alignment to a reference genome than by de novo read assembly. As such, the early use of NGS sparked the development of numerous read alignment algorithms, each using one of several different principles to match each short read to an existing genome assembly. Following read alignment to a reference genome, there was an ensuing need to identify variants—nucleotide sequence differences between the sequenced genome and the reference—and to subsequently interpret these changes in terms of their impact on protein coding genes, regulatory regions, and other sequence data features. Here, the most straightforward variant calling algorithms to develop were those capable of detecting single-nucleotide variants (SNVs), also known as substitutions or point mutations. Less straightforward was the detection of insertion or deletion events ("in-

dels"), in which one or more nucleotides were either added or deleted in comparison to the reference genome. These variants remain quite difficult to detect because of difficulties in aligning the sequence read with indels to the reference genome, although the accuracy of their detection has been greatly enhanced by the increased read lengths of NGS platforms over time. Beyond read length, the development of paired-end read approaches also has enhanced our ability to detect structural variants, especially for complex repetitive genomes. In paired-end sequencing, a second set of SBS reads is generated by priming the opposite end of each library fragment with an adaptor-specific oligonucleotide, thereby generating concordant X–Y data from the opposite end of each library fragment. These read pairs are then mapped to the genome, using information about their predicted separation (based on insert size for the library) and read orientation. Large-scale genomic alterations that represent sequence duplications or amplifications, large deletions, or structural alterations (translocations, inversions) continue to represent computational challenges to NGS data interpretation although, as for indels, their accuracy and false-positive rates have improved somewhat with longer read lengths and read pairs that provide more accurate read mapping and read directionality. Here, reads that group outside of the predicted fragment length separation or with different orientations than predicted, or both, are subject to secondary analysis that may predict large-scale structural changes in comparison to the reference genome. Typical analytical pipelines that have the capability to detect these more challenging alterations combine multiple algorithms to examine the aligned read data and produce a consensus report—those alterations that are identified by multiple different algorithms and supported by multiple read pairs—which are then subject to manual scrutiny and often secondary validation to support or refute them. As described in the final section of this article, read alignment is just the starting point for many types of NGS analyses, dictated entirely by the upfront preparatory assays that are used to generate the sequencing library. As such, NGS experiments link together experimental design with data analysis in intimate and inextricable ways that has helped to shape the broad-based use of this fundamental tool.

Another type of analysis of NGS data uses the sequencer as a counting machine. In RNA-Seq, sequences from different transcripts can be counted and their relative number compared. This allows a relatively accurate description of the transcript abundance profile in a given tissue or cell source. With the large number of reads that can be obtained from next-gen sequencers the relative frequency of transcripts can be obtained over a very high dynamic range. Any biological assay that can be converted into sequence fragment counting can be detected in this way.

SINGLE-MOLECULE SEQUENCING (SMS) TECHNOLOGIES

Basics of SMS

Massively parallel sequencing instruments using SBS approaches use an enzymatic on-surface amplification of each library fragment before the initiation of the sequencing reaction, to produce sufficient signal for detection by the instrument. This intermediate amplification step, while necessary, introduces a variety of artifacts. One type of artifact is caused by the inherent error rate of polymerases, some of which can masquerade as true variants if they occur sufficiently early in the fragment amplification process. The other artifact is the result of nucleotide composition of the fragment, wherein fragments with high or low G + C composition are less efficiently amplified compared with more equal ratio composition fragments. This step also somewhat limits the length of SBS library inserts because of a desire to make the amplification rapid, efficient, and consistent across all library fragments. In addition, as mentioned above, as the clusters of molecules are sequenced, some molecules in each cluster misincorporate the added base on each cycle of nucleotide addition. This has the effect of increasing the noise at each nucleotide addition, which ultimately results in the noise being too high to allow further sequencing. This too limits read lengths in SBS.

As a result, SMS methods and instrumentation have been pursued to obviate library fragment amplification steps and associated artifacts, which in turn permits increased library fragment sizes and read lengths.

Although the advantages of SMS are obvious, the difficulty in overcoming the signal-to-noise aspects of collecting data from a single molecule are considerable, regardless of the approach. As a result, SMS reads typically have a higher error rate than SBS reads and have required both higher sequence data coverage and novel computational approaches to analyze the resulting data. Despite these challenges, two SMS devices have achieved commercial status, each using highly unique approaches to sequencing detection.

Single-Molecule Fluorescent Sequencing

The initial SMS device to achieve commercial release was the Pacific Biosciences instrument, which emerged from a proof-of-principle instrumentation concept published by Watt Webb's group at Cornell University (Levene et al. 2003). This instrument used patterned nanofabricated chambers on a silicon surface, called zero-mode waveguides (ZMWs), to isolate individual polymerase enzymes coupled to primed DNA template molecules. On supplying these polymerases with physiological concentrations of fluorescent labeled nucleotides, the copying of each template can progress. Here, the ZMW provides a mechanism to focus the detection apparatus onto the polymerase, effectively permitting the instrument optics to collect data in real time as the polymerase copies the template. The resulting "movies" represent optical data collected at the active site of each polymerase, in which an incoming nucleotide is detected successfully by virtue of its increased dwell time in the active site during incorporation. On its addition to the growing synthesized strand, each nucleotide loses its fluorescent label, which is covalently attached onto the phosphate portion of the molecule and diffuses out of focus. Incorrect template-matching nucleotides can enter the active site but typically have insufficient dwell time of detection of their attached fluor to be measured as a nucleotide incorporation. Sources of error for this sequencing method arise when a correct nucleotide is incorporated too quickly for sufficient detection time by the optics, or when multiple nucleotides are incorporated in quick succession without resolution by the optics of each incorporation (e.g., mononucleotide stretches), both of which result in deletions in the sequence read. A deletion can also result from incorporation of an unlabeled ("dark") nucleotide, although these are rare. Another source of error occurs when the incorrect nucleotide dwells too long in the active site and is detected by the optics yet is not actually incorporated by the polymerase, resulting in an insertion error.

The Pacific Biosciences platform has increased average read length and accuracy over time by virtue of a variety of enzymatic, nucleic acid chemistry, and optical detector sensitivity improvements. Current per-read error rate is ~10%, although these are random errors (e.g., not systematic or predictable), which means that the consensus error rate improves as coverage depth increases. Improvements to the method and instrument have cumulatively allowed for increasing library fragment lengths, commensurate with increased optical imaging duration for data collection. This, in turn, has placed renewed emphasis on techniques to produce fragment lengths in excess of 10,000 bp for library construction. Throughput per instrument run also has been impacted by increasing the number of ZMWs from which data can be collected, in combination with the aforementioned increases in read length. It is now possible to have average read lengths in excess of 10,000 bases with this instrument compared with 150–250 base maximum read lengths with Illumina sequencers.

Single-Molecule Nanopore-Based Sequencing

A second approach to SMS, commercialized by Oxford Nanopore Technologies, uses translocation of single DNA strands through a surface-positioned nanopore grid to generate sequencing data by sensing the changes in electrical conductance during nucleotide translocation of the DNA strand through the pore. Each

nanopore has a detector positioned adjacent to it that records these conductance changes over time of fragment translocation through the pore. The nanopore sequencing library is produced much like other NGS libraries, and uses adaptor sequences that have a platform-specific composition, basecalling model that is built on known sequence context-conductance correlation data, and the resulting sequence read is interpreted computationally from the best fit to the model. Like the Pacific Biosciences data, the single-molecule nanopore data is subject to signal-to-noise constraints and has a correspondingly high error rate when compared with SBS data. This error rate also has improved over time and development of improved basecalling models, improved polymerase properties, and modifications to the nanopore that slow the traverse of the molecules through the nanopores, permit more data richness that correspondingly provides a better fit to the model. Unlike other platforms, many errors are not random, being sequence context dependent, resulting in error not as readily resolved with increasing coverage in these regions. Similarly, read lengths have improved over time as well, shifting focus to the development of library construction approaches that produce longer fragments. Unlike the Pacific Biosciences technology and all of the previously discussed types of SBS platforms, the Oxford Nanopore sequencing method does not require a polymerase or labeled nucleotides to conduct sequence data generation. At its essence, the nanopore is the only reagent needed for the platform to sequence the library fragments.

Analysis of Combined Short- (SBS) and Long-Read (SMS) Data

Long-read lengths offer significant advantages in genome assembly, yet, initially, the high error rates obtained from SMS platforms made direct assembly of long-read SMS data difficult to achieve. To address these difficulties, several groups have explored different approaches that combined SBS and SMS data to achieve high quality and contiguity in genome assemblies. Either the long-read data from SMS platforms were assembled as a scaffold against which short-read SBS data could be aligned, or the SBS data were aligned to the longer reads to improve overall accuracy and the resulting highly accurate long reads were assembled. Despite these innovative approaches, the largest genomes routinely assembled with these approaches were bacterial genomes (Ribeiro et al. 2012). The benefit of these early innovations to long-read SMS data analysis was that they fueled the development of sequence assembly approaches suitable for read data with a random error model, and as the sequence data quality and read lengths of SMS reads improved, these algorithms became suitable for long-read assembly without the need for SBS error correction.

Analysis of Long-Read Single-Molecule Data

The combination of increased read lengths from SMS platforms, improved error rates, and optimized assembly algorithms has now resulted in the singular use of SMS data for genome assembly. There are distinct advantages to this approach that may be more or less important, depending on the specific application. In particular, long reads offer long-range contiguity and, hence, for diploid organism haplotypes are generated that span long distances along a chromosome. Similarly, complex repetitive sequences may be elucidated from long-read data, enabling complicated genomic regions to be accurately and completely sequenced even where highly similar tandem duplications exist (also known as segmental duplications) (Huddleston et al. 2014). Indeed, as the capability for long-read assembly emerged, sequencing and assembly of a human genome using Pacific Biosciences SMS data was compared with the same genome sequenced and aligned using Illumina SBS data (Chaisson et al. 2015). This comparison highlighted the advantages of long-read SMS assembly to provide contiguity, haplotype resolution, and to deconvolute the sequence data in complex areas of the human genome that are simply not resolvable by short-read data. One might then raise the obvious question of why more human resequencing efforts are not being pursued with SMS data generation. The primary reason at this writing for continued human

(and other complex organisms) genome sequencing using SBS over SMS is simply one of cost and throughput, wherein SBS offers clear advantages for both. Similarly, the short-read libraries needed for SBS are more straightforward to generate, are more readily automated, and require significantly less input DNA, making them suitable for clinical samples that often provide limited amounts of DNA.

The use of SMS sequencing, either alone or in combination with SBS data, continues to be an active research area. Human genomes can reliably assemble into considerably more contiguous genomes than is possible with SBS alone. Recently, even larger genomes, such as the 18-Gb wheat genome have been assembled with combined SMS and SBS data (Zimin et al. 2017). Applications of NGS in biomedical research and clinical diagnostics include (1) DNA-based applications in research, (2) RNA-based applications in research, (3) combinatorial DNA–RNA interaction applications, and (4) clinical applications of NGS.

REFERENCES

Barnes WM. 1978. DNA sequencing by partial ribosubstitution. *J Mol Biol* **119:** 83–99. doi:10.1016/0022-2836(78)90271-1

Chaisson MJ, Huddleston J, Dennis MY, Sudmant PH, Malig M, Hormozdiari F, Antonacci F, Surti U, Sandstrom R, Boitano M, et al. 2015. Resolving the complexity of the human genome using single-molecule sequencing. *Nature* **517:** 608–611. doi:10.1038/nature13907

Ewing B, Green P. 1998. Base-calling of automated sequencer traces using *phred*. II: Error probabilities. *Genome Res* **8:** 186–194. doi:10.1101/gr.8.3.186

Ewing B, Hiller L, Wendl MC, Green P. 1998. Base-calling of automated sequencer traces using *phred*. I: Accuracy assessment. *Genome Res* **8:** 175–185. doi:10.1101/gr.8.3.175

Gordon D, Abajian C, Green P. 1998. *Consed*: A graphical tool for sequence finishing. *Genome Res* **8:** 195–202. doi:10.1101/gr.8.3.195

Huddleston J, Ranade S, Malig M, Antonacci F, Chaisson M, Hon L, Sudmant PH, Graves TA, Alkan C, Dennis MY, et al. 2014. Reconstructing complex regions of genomes using long-read sequencing technology. *Genome Res* **24:** 688–696. doi:10.1101/gr.168450.113

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, et al. 2001. Initial sequencing and analysis of the human genome. *Nature* **409:** 860–921. doi:10.1038/35057062

Levene MJ, Korlach J, Turner SW, Foquet M, Craighead HG, Webb WW. 2003. Zero-mode waveguides for single-molecule analysis at high concentrations. *Science* **299:** 682–686. doi:10.1126/science.1079700

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. doi:10.1038/nature03959

Maxam AM, Gilbert W. 1977. A new method for sequencing DNA. *Proc Natl Acad Sci* **74:** 560–564. doi:10.1073/pnas.74.2.560

Ribeiro FJ, Przybylski D, Yin S, Sharpe T, Gnerre S, Abouelleil A, Berlin AM, Montmayeur A, Shea TP, Walker BJ, et al. 2012. Finished bacterial genomes from shotgun sequence data. *Genome Res* **22:** 2270–2277. doi:10.1101/gr.141515.112

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. doi:10.1016/0022-2836(75)90213-2

Sanger F, Nicklen S, Coulson AR. 1977. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci* **74:** 5463–5467. doi:10.1073/pnas.74.12.5463

Sanger F, Coulson AR, Hong GF, Hill DF, Petersen GB. 1982. Nucleotide sequence of bacteriophage λDNA. *J Mol Biol* **162:** 729–773. doi:10.1016/0022-2836(82)90546-0

Smith LM, Sanders JZ, Kaiser RJ, Hughes P, Dodd C, Connell CR, Heiner C, Kent SB, Hood LE. 1986. Fluorescence detection in automated DNA sequence analysis. *Nature* **321:** 674–679. doi:10.1038/321674a0

Staden R. 1979. A strategy of DNA sequencing employing computer programs. *Nucleic Acids Res* **6:** 2601–2610. doi:10.1093/nar/6.7.2601

Staden R. 1984. Computer methods to aid the determination and analysis of DNA sequences. *Biochem Soc Trans* **12:** 1005–1008. doi:10.1042/bst0121005

Staden R, Beal KF, Bonfield JK. 2000. The Staden Package, 1998. *Methods Mol Biol* **132:** 115–130.

Zimin AV, Pulu D, Hall R, Kingan S, Clavijo BJ, Salzberg SL. 2017. The first near-complete assembly of the hexaploid bread wheat genome, *Triticum aestivum*. *Gigascience* **6:** 1–7.