A defining decade in DNA sequencing

John D McPherson

A revolution in DNA sequencing technology has enabled new insights from thousands of genomes sequenced across taxa.

© 2014 Nature America, Inc. All rights reserved.

The tenth anniversary of *Nature Methods* offers an excellent occasion to reflect on protocols and technologies that are rapidly shaping research methodology. A major advance has undoubtedly been the massive increase in DNA and RNA sequencing capabilities that fall under the general term of next-generation sequencing (NGS). The death in 2013 of Frederick Sanger, who pioneered methods for deciphering amino acid and nucleic acid sequences, for which he was awarded two Nobel Prizes (1958 and 1980), also affords the opportunity to look back at recent events and the extraordinary progress that has come from the sequencing metamorphosis. Many new sequencing platforms have matured while others have failed to gain a market share, but the result has been an extraordinary increase in sequencing capabilities (**Fig. 1**).

For many years, Sanger DNA sequencing—a method that utilizes dideoxynucleotide analogs to halt template elongation, resulting in a ladder of fragments separable on a gel matrix—dominated the DNA analysis field. Its dominance was ensured by the introduction of fluorescent markers, which enabled the automation of sequence data collection. Incremental improvements on this basic concept produced a fully automated platform capable of generating 500,000 bases per day, enabling the first sequencing of the human genome in less than a decade. The Sanger sequencing instrumentation space was dominated then and now by a single vendor—Applied Biosystems, now under Life Technologies as part of Thermo Fisher Scientific—and these platforms are still the mainstay in clinical diagnostic DNA sequencing settings.

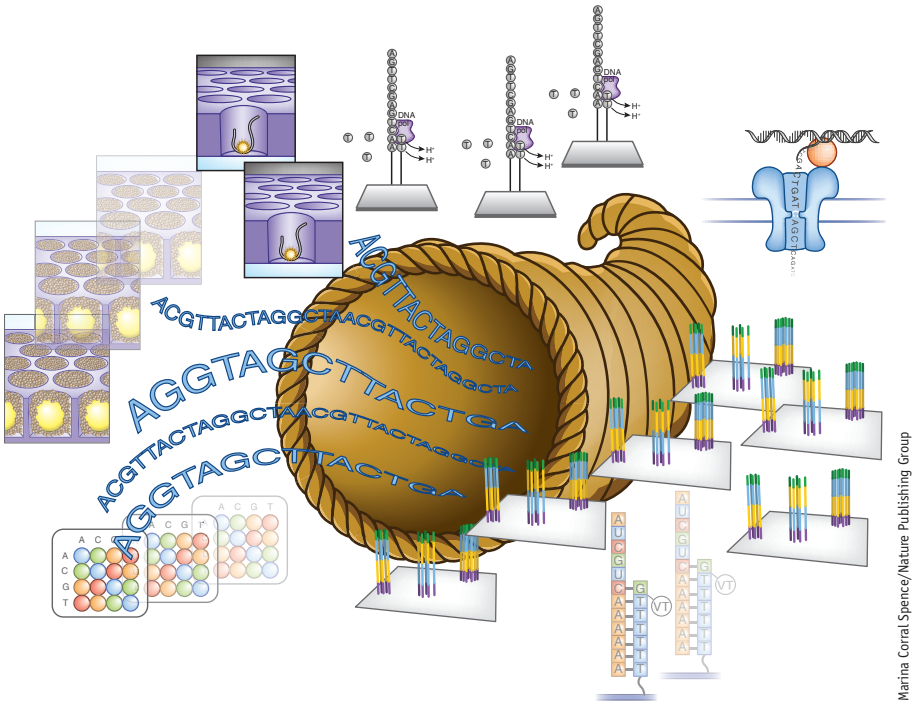
An often-overlooked milestone in nucleic acid sequencing was the massively parallel signature sequencing (MPSS) system by Lynx Therapeutics, which pre-dated current NGS technologies by 5 years[1]. Its relatively high cost for data generation prevented broad adoption, but it was the first non-Sanger high-throughput platform and gave a glimpse of the coming NGS era. The DNA sequencing landscape was then dramatically altered, 4 years after the release of the first human genome, with a landmark publication detailing a pyrosequencing platform capable of generating 25 million bases in a single 4-hour run cycle[2]. The 454 Genome Sequencer (GS20) based on this technology was the first commercial NGS system designed for individual laboratory use (454 is now part of Roche Diagnostics). In the nearly 10 years since this launch, the competition within the NGS platform space has been intense.

The next major player was Solexa, which employed reversible dideoxy terminators to add a single base at a time (sequencing-by-synthesis (SBS)) to amplified clusters of single DNA molecules. Solexa acquired Lynx Biotherapeutics while bringing their first platform, the Genome Analyzer (GA), to market. The GA produced a remarkable number of sequence reads, generating 1 gigabase of data, but the very short read length limited its utility in complex genomes. Illumina acquired Solexa to further develop the GA, and shortly thereafter, Applied Biosystems entered the NGS market with the SOLiD system—a ligation-based chemistry platform. The arms race between Applied Biosystems (which later merged with Invitrogen to form Life Technologies) and Illumina began in earnest with read length and total number of reads largely defining the weaponry. This war has been decisively won by Illumina, which dominates the sequencing space, but not without competition. Achievement of the $1,000 genome has long been seen as the ultimate victory, which Illumina claims to have reached.

The DNA sequencing platform space has begun to diversify, with the arrival of new entries into the field every year or two. Platforms capable of sequencing single unamplified molecules emerged. Helicos Biosciences introduced the Heliscope, a powerful DNA microscope, promising unprecedented sequencing speed without bias due to template amplification. This was truly a remarkable achievement, but the platform ultimately could not compete in an atmosphere of increasing read lengths and lowering costs and could not distinguish itself for applications justifying its high instrumentation cost. Helicos Biosciences declared bankruptcy in 2012, just 9 years after its founding.

Pacific Biosciences (PacBio), founded in 2004, launched their single-molecule sequencing platform through an early-access program in 2010 to mixed reviews. Extraordinary read lengths now exceed 4,000 bases, but this comes with a high error rate of ~15%. The consensus accuracy of overlapping reads is high, but tools for handling these long, imprecise reads were slow to develop, and with read numbers measured on the order of hundreds of thousands rather than milllions, its ready application to large genomes is stunted. Hybrid sequencing approaches combining these long reads with plentiful and accurate short reads have emerged. These early experiences in launching platforms have led many companies to develop new platforms

John D. McPherson is in the Genome Technologies Program at the Ontario Institute for Cancer Research, Toronto, Canada.
e-mail: john.mcpherson@oicr.on.ca



Marina Corral Spence/Nature Publishing Group

Figure 1 | Sequencing technologies are evolving, with many platforms now reaching maturity as a few others have failed or are being phased out. The future face of sequencing is unclear, but a bountiful sequencing capacity is assured for the future.

more quietly. This stealth mode of product development is now generally the hallmark of the NGS industry, as early performance metrics are ruthlessly scrutinized and compared to those of mature platforms by an impatient research community.

An important concept for NGS has been the democratization of sequencing—moving large-scale sequencing out of the few genome centers established largely for the Human Genome Project. This was fully embraced by Ion Torrent Systems (later acquired by Life Technologies), which introduced the Personal Genome Machine (PGM), a true benchtop sequencer that was affordable to most labs, in 2010. The PGM uses semiconductor detection of protons released as nucleotides are incorporated in a growing DNA strand, obviating the need for costly optics. The PGM, and its successor, the Proton, were released with great fanfare declaring sequencing available to all, with the latter said to be on the cusp of attaining the coveted $1,000 human genome. Although the platform has delivered on low-cost entry into NGS, it has fallen short of its benchmarks for per-run sequencing output. Nevertheless, it continues to make application headway, with throughput gains likely to be possible. Illumina also entered this lower-cost barrier-to-entry market segment with the MiSeq. Both platforms are gaining traction in the clinical environment, filling a need for short-turnaround, lower-throughput rapid DNA diagnostics. A notable change in this NGS area is the announcement that Roche will close down its 454 operations in mid-2016. One can only speculate about the reason, but in this highly competitive field with increasing throughput and decreasing costs, a platform must evolve along with these needs or perish.

The past decade has seen a transformation, fueled by the availability of NGS platforms, in applications based on generating DNA and RNA sequence. There are more than 50 multiletter-acronym or 'MLA-seq' applications for interrogating a wide variety of genome characteristics. One of the first was RNA-seq, for transcript quantification and characterization. This quickly eclipsed microarrays for its superior dynamic range, sensitivity and ability to discover transcripts not included in the microarray probe set. RNA-seq has greatly expanded the view of the transcriptome, underpinning the importance of non–protein coding transcripts. Chromatin immunoprecipitation sequencing (ChIP-seq) allowed the genome-wide positioning of protein-DNA interactions, including transcription factor binding sites and histone modification patterns related to gene expression. Epigenetic DNA modifications such as methylation have been assayed using whole-genome sequencing of bisulfite-treated DNA (WGBS-seq). These and other methods have led to comprehensive catalogs of functional genomic elements, best exemplified by the 2012 release of 30 manuscripts describing more than 1,600 annotation data sets for the Encyclopedia of DNA Elements (ENCODE), a project started shortly after the completion of the human genome to provide functional insight into the landmark DNA sequence framework. More improvements are needed to these workflows to make these functional element assays as robust and efficient as the determination of single base substitutions.

A potential shift in the democratization of sequencing has been to render it as an affordable service rather than moving platforms into individual laboratories. Human genome sequencing costs have fallen well below $5,000, driven by competition between companies such as Illumina and Complete Genomics. The most recent foray in this area has been Illumina's HiSeq X Ten platform, which delivers the $1,000 human genome but with an appetite for ~20,000 human genomes per year. This is an outstanding achievement, empowering future large-scale, population-based human genetic and medical research. The Cancer Genome Atlas and the International Cancer Genome Consortium have already collectively sequenced thousands of tumor genomes, lending insights into the disease's underlying mechanisms but also indicating that its extreme heterogeneity will require much more sequencing. The characterization of microbiomes of many environments has provided a glimpse of the organismal diversity surrounding and within us and its association with both healthy and disease states. The full realization of DNA-based diagnostics and guided therapies will require the coming sequencing scales and diminished costs.

Applications beyond the human genome, with varying needs of input and analysis, still call for more flexible systems. Although it is dominated by Illumina, the NGS industry is still shifting and is subject to new platform technologies, enabling diverse applications and leaving room for niche markets. The horizon is populated with protein and semiconductor nanopore

© 2014 Nature America, Inc. All rights reserved.

technologies promising extraordinarily long read lengths. Oxford Nanopore has been the first to launch such a system, and its read length, accuracy and read number metrics are emerging. The technology looks promising but does not yet herald another rapid transformation of the face of sequencing by a disruptive platform. The course of sequencing technologies is difficult to predict, but a shift from template-guided DNA replication methods that use optical measurements to more direct interrogation of native molecules is likely.

The huge capacity for sequencing shifts the bottlenecks for genome analysis. The magnitude of the coming capacity will require unimaginable numbers of DNA and RNA samples, collected properly and with the appropriate consent. Extraction and storage of analytes from a large number and variety of tissue types and specimens presents considerable challenges. The large amounts of data generated per sample already strain storage infrastructures and analysis pipelines. Much improvement is needed to speed sequence alignment to reference genomes and to derive a fully accurate accounting of nucleotide substitutions and larger genomic alterations. Genome assembly is preferable to reference alignment but is not yet tractable for large genomes. Reaping the full benefit of genome sequencing will require vast improvements in predicting the functional significance of coding variants and noncoding elements. The variety of sequence-based genome-feature analyses possible through MLA-seq methods may further complicate this picture as the importance of these features in disease states is revealed.

The past decade has seen a remarkable change in DNA sequencing technology that can be characterized truly as a revolution. The evolving platforms have been well described in reviews spanning this same time period[3–6]. The success of NGS adoption is evident through its prolific use in publications (>6,300 in PubMed, in a search for publications with 'next generation sequencing' in their title or abstract). The reduced cost and increased throughputs have enabled very deep sequencing of genomes across taxa. Much like the evolution of the optical microscope, this change has come with an ever-increasing understanding of the depth of this research space, which largely offers more questions than concrete answers. The coming decade holds great promise, but as in the one past, the full magnitude of the coming changes to DNA sequencing are difficult to anticipate. If the current trajectory continues, the question may shift from one of whether your genome will be sequenced to one of when. The limitations lie in the ability to further disseminate cost-effective sequencing, interpret the vast quantities of data to make effective use of decoding DNA and for people to give consent for targeted gene or whole-genome deciphering—a decision future generations will certainly face.

COMPETING FINANCIAL INTERESTS
The author declares no competing financial interests.

1. Brenner, S. *et al. Nat. Biotechnol.* **18**, 630–634 (2000).
2. Margulies, M. *et al. Nature* **437**, 376–380 (2005).
3. Metzker, M.L. *Genome Res.* **15**, 1767–1776 (2005).
4. Mardis, E.R. *Annu. Rev. Genomics Hum. Genet.* **9**, 387–402 (2008).
5. Shendure, J. & Ji, H. *Nat. Biotechnol.* **26**, 1135–1145 (2008).
6. Metzker, M.L. *Nat. Rev. Genet.* **11**, 31–46 (2010).

© 2014 Nature America, Inc. All rights reserved.