Next-Generation Sequencing Platforms

Elaine R. Mardis

The Genome Institute at Washington University School of Medicine, St. Louis, Missouri 63108; email: emardis@wustl.edu

Annu. Rev. Anal. Chem. 2013. 6:287–303

The *Annual Review of Analytical Chemistry* is online at anchem.annualreviews.org

This article's doi: 10.1146/annurev-anchem-062012-092628

Keywords

massively parallel sequencing, next-generation sequencing, reversible dye terminators, sequencing by synthesis, single-molecule sequencing, genomics

Abstract

Automated DNA sequencing instruments embody an elegant interplay among chemistry, engineering, software, and molecular biology and have built upon Sanger's founding discovery of dideoxynucleotide sequencing to perform once-unfathomable tasks. Combined with innovative physical mapping approaches that helped to establish long-range relationships between cloned stretches of genomic DNA, fluorescent DNA sequencers produced reference genome sequences for model organisms and for the reference human genome. New types of sequencing instruments that permit amazing acceleration of data-collection rates for DNA sequencing have been developed. The ability to generate genome-scale data sets is now transforming the nature of biological inquiry. Here, I provide an historical perspective of the field, focusing on the fundamental developments that predated the advent of next-generation sequencing instruments and providing information about how these instruments work, their application to biological research, and the newest types of sequencers that can extract data from single DNA

1. INTRODUCTION

Automated DNA sequencing instruments embody an elegant interplay among chemistry, engineering, software, and molecular biology and have built upon Sanger's founding discovery of dideoxynucleotide sequencing to perform once-unfathomable tasks. Combined with innovative physical mapping approaches that helped to establish long-range relationships between cloned stretches of genomic DNA, fluorescent DNA sequencers have been used to produce reference genome sequences for model organisms (*Escherichia coli*, *Drosophila melanogaster*, *Caenorhabditis elegans*, *Mus musculus*, *Arabidopsis thaliana*, *Zea mays*) and for the reference human genome. Since 2005, however, new types of sequencing instruments that permit amazing acceleration of data-collection rates for DNA sequencing have been introduced by commercial manufacturers. For example, single instruments can generate data to decipher an entire human genome within only 2 weeks. Indeed, we anticipate instruments that will further accelerate this whole-genome sequencing data–production timeline to days or hours in the near future. The ability to generate genome-scale data sets is now transforming the nature of biological inquiry, and the resulting increase in our understanding of biology will probably be extraordinary. In this review, I provide an historical perspective of the field, focusing on the fundamental developments that predated the advent of next-generation sequencing instruments, providing information about how massively parallel instruments work and their application to biological research, and finally discussing the newest types of sequencers that are capable of extracting sequence data from single DNA molecules.

2. A BRIEF HISTORY OF DNA SEQUENCING

DNA sequencing and its manifest discipline, known as genomics, are relatively new areas of endeavor. They are the result of combining molecular biology with nucleotide chemistry, both of which blossomed as scientific disciplines in the 1950s. Dr. Frederick Sanger's laboratory at the Medical Research Council (MRC) in Cambridge, United Kingdom, began research to devise a method of DNA sequencing in the early 1970s (1–3) after having first published methods for RNA sequencing in the late 1960s (4–6). Sanger et al.'s (7) seminal 1977 publication describes a method for essentially tricking DNA polymerase into incorporating nucleotides with a slight chemical modification—the exchange of the 3′ hydroxyl group needed for chain elongation with a hydrogen atom that is functionally unable to participate in the reaction with the incoming nucleotide to extend the synthesized strand. Mixing proportions of the four native deoxynucleotides with one of four of their analogs, termed dideoxynucleotides, yields a collection of nucleotide-specific terminated fragments for each of the four bases (**Figure 1**). The fragments resulting from these reactions were separated by size on thin slab polyacrylamide gels; the A, C, G, and T reactions were performed for each template run in adjacent lanes. The fragment positions were identified by virtue of ^{32}P, which was supplied in the reaction as labeled dATP molecules. When dried and exposed to X-ray film, the gel-separated fragments were visualized and subsequently read from the exposed film from bottom to top (shortest to longest fragments) by the naked eye. Thus, a long and labor-intensive process was completed, and the sequencing data for the DNA of interest were in hand and ready for assembly, translation to amino acid sequence, or other types of analysis.

Sequencing by radiolabeled methods underwent numerous improvements following its invention until the mid 1980s. These improvements included the invention of DNA synthesis chemistry (8, 9) and, ultimately, of DNA synthesizers that can be used to make oligonucleotide primers for the sequencing reaction (providing a 3′-OH for extension); improved enzymes from the original

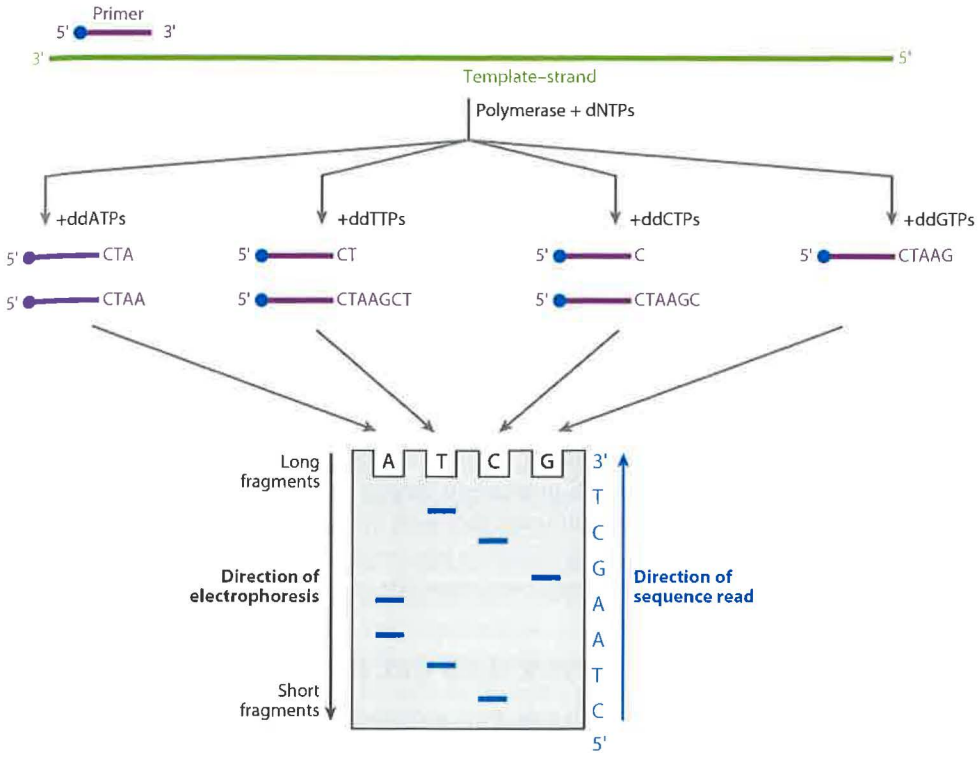

Figure 1

Sanger sequencing.

use of ^{35}S- in place of ^{32}P-dATP for radiolabeling (sharper banding and hence longer read lengths); and the use of thinner and/or longer polyacrylamide gels (improved separation and longer read lengths), among others. Although there were attempts at automating various steps of the process, notably the automated pipetting of sequencing reactions and the automated reading of the autoradiograph banding patterns, most improvements were not sufficient to make this sequencing approach truly scalable to high-throughput needs.

3. IMPACT OF FLUORESCENCE LABELING

A significant change in the scalability of DNA sequencing was introduced in 1986, when Applied Biosystems, Inc. (ABI), commercialized a fluorescent DNA sequencing instrument that had been invented in Leroy Hood's laboratory at the California Institute of Technology (12). In replacing the use of radiolabeled dATP with reactions primed by fluorescently labeled primers (different fluor for each nucleotide reaction), the laborious processes of gel drying, X-ray film exposure and developing, reading autoradiographs, and performing hand entry of the resulting sequences were eliminated. In this instrument, a raster scanning laser beam crossed the surface of the gel plates to provide an excitation wavelength for the differentially labeled fluorescent primers to be detected during the electrophoretic separation of fragments. Thus, significant manual effort and several sources of error were eliminated. By use of the initial versions of this instrument

laboratories used newly available automated pipetting stations to decrease the effort and error rate of the upstream sequencing reaction pipetting steps (13). During this time, investigators made additional improvements to sequencing enzymology and processes, including the ability to perform cycled sequencing reactions catalyzed by thermostable sequencing polymerases (14) that were patterned after the polymerase chain reaction (PCR), which was first described in 1988 by Mullis and colleagues (15). By incorporating linear (cycled) amplification into the sequencing reaction, one could begin with significantly lower input template DNA and hence could produce uniform results across a range of DNA yields (from automated isolation methods in multiwell plates, for example). Improvements to chemistry were also important, as fluorescent dye–labeled dideoxynucleotides (known as terminators) were introduced (16). Because the terminating nucleotide was identified by its attached fluor, all four reactions could be combined into a single reaction, greatly decreasing the cost of reagents and the input DNA requirements. Finally, the per run throughput of the sequencers increased during this time (17), ultimately permitting 96 samples to be loaded on one gel. These technological breakthroughs combined to make 96-well and ultimately 384-well sequencing reactions a major contributor to scalability. These high-throughput slab gel fluorescence instruments largely contributed to the sequencing of several model organism genomes, and although they were impressive in their capacity to produce data, they still contained several manual and hence labor-intensive and error-prone steps. These limitations largely centered around casting polyacrylamide gels and loading samples by hand.

4. IMPACT OF CAPILLARY OVER SLAB GEL ELECTROPHORESIS

The rate-limiting manual steps in slab gels were addressed in 1999 with the introduction of capillary sequencing instruments, first the MegaBACE™ sequencer from Molecular Dynamics (18) and then the ABI PRISM® 3700. These instruments solved the slab gel problem by directly injecting a polymeric separation matrix into capillaries that provided single-nucleotide resolution. Samples, by definition, could also be loaded directly from the microtiter plate to the capillaries for separation by use of electrical current pulses through a process known as electrokinetic injection. Following the separation and detection of reaction products, the polymer matrix was replaced by pumping in new matrix. Thus, these instruments eliminated an entire series of rate-limiting steps. Downstream activities were further simplified because the capillaries were fixed in their positions, so there was no need for tracking lanes on the slab gel image, and subsequent data extraction and base-calling were much faster and more accurate. Lastly, the run times were greatly accelerated due to the rapid heat dissipation of the capillaries over thick glass plates. The ABI PRISM 3700 instruments and a later upgrade (ABI 3730) were principal data-generating instruments for the human and mouse genome projects, among others. Their scalability and ease of use came at a crucial time, when large-scale robotics to perform DNA extraction and sequencing were available in specialized facilities for the clone-based front end of the process.

Indeed, these reference genomes that were produced for major model organisms, human and plant, provided not only a fundamental advance for biological studies in these organisms but also the basis for the utility of next-generation sequencing instruments. Next-generation sequencing is described in the next section.

5. GENERAL PRINCIPLES OF NEXT-GENERATION SEQUENCING

Beginning in 2005, the traditional Sanger-based approach to DNA sequencing has experienced revolutionary changes (19, 20). The previous "top-down" approach involved characterizing large

subclones that were assembled and finished to recapitulate each originating, larger clone (21). The sequences of the larger clones were then stitched together at their overlapped ends to reconstruct entire chromosomes (with small gaps). By contrast, next-generation sequencing instruments do not require a cloning step per se. Rather, the DNA to be sequenced is used to construct a library of fragments that have synthetic DNAs (adapters) added covalently to each fragment end by use of DNA ligase. These adapters are universal sequences, specific to each platform, that can be used to polymerase-amplify the library fragments during specific steps of the process. Another difference is that next-generation sequencing does not require performing sequencing reactions in microtiter plate wells. Rather, the library fragments are amplified in situ on a solid surface, either a bead or a flat glass microfluidic channel that is covalently derivatized with adapter sequences that are complementary to those on the library fragments. This amplification is digital in nature; in other words, each amplified fragment yields a single focus (a bead- or surface-borne cluster of amplified DNA, all of which originated from a single fragment). Amplification is required to provide sufficient signal from each of the DNA sequencing reaction steps that determine the sequencing data for that library fragment. The scale and throughput of next-generation sequencing are often referred to as massively parallel, which is an appropriate descriptor for the process that follows fragment amplification to yield sequencing data. In Sanger sequencing, the reaction that produces the nested fragment set is distinct from the process that separates and detects the fragments by size to produce a linear sequence of bases. In massively parallel sequencing, the process is a stepwise reaction series that consists of (*a*) a nucleotide addition step, (*b*) a detection step that determines the identity of the incorporated nucleotides on each fragment focus being sequenced, and (*c*) a wash step that may include chemistry to remove fluorescent labels or blocking groups. In essence, next-generation sequencing instruments conduct sequencing and detection simultaneously rather than as distinct processes, one of which is completed before the other takes place. Moreover, these steps are performed in a format that allows hundreds of thousands to billions of reaction foci to be sequenced during each instrument run and, hence, at a capacity per instrument that can produce enormous data sets.

One final difference between Sanger sequencing data and next-generation sequencing data is the read length, or the number of nucleotides obtained from each fragment being sequenced. In Sanger sequencing, the read length was determined largely by a combination of gel-related factors, such as the percentage of polyacrylamide, the electrophoresis conditions, the time of separation, and the length and thickness of the gel. In next-generation sequencing, the read length is a function of the signal-to-noise ratio. Because the sources of noise differ according to the technology, specifics are described for each type of sequencing below. However, the major impact of the signal-to-noise ratio is to limit the read length from all next-generation sequencing instruments, all of which produce shorter reads than does Sanger sequencing.

Shorter read lengths, in turn, are a differentiation point because, although short reads can be assembled as are traditional Sanger reads, based on shared sequence, the lower extent of shared sequence (due to read length) limits the ability to assemble these reads, so the overall length of contiguous sequence that can be assembled is limited. This limitation is exacerbated by genome size and complexity (e.g., repetitive content and gene families), so genomes such as that of the human (3 Gb and ~48% repetitive content) cannot be reassembled from the component reads of a whole-genome shotgun of next-generation sequencing data. Rather, because a high-quality reference genome exists for many model organisms and for humans, sequence read alignment is a more practical approach to sequencing data analysis from next-generation read lengths. Specific algorithms to approach short read alignment have been devised; they provide a score-based metric indicative of that sequence's best fit in the genome, whereby sequences that contain mostly or

DOCKET ALARM

Explore Litigation Insights

Docket Alarm provides insights to develop a more informed litigation strategy and the peace of mind of knowing you're on top of things.

Real-Time Litigation Alerts

Keep your litigation team up-to-date with **real-time alerts** and advanced team management tools built for the enterprise, all while greatly reducing PACER spend.

Our comprehensive service means we can handle Federal, State, and Administrative courts across the country.

Advanced Docket Research

With over 230 million records, Docket Alarm's cloud-native docket research platform finds what other services can't. Coverage includes Federal, State, plus PTAB, TTAB, ITC and NLRB decisions, all in one place.

Identify arguments that have been successful in the past with full text, pinpoint searching. Link to case law cited within any court document via Fastcase.

Analytics At Your Fingertips

Learn what happened the last time a particular judge, opposing counsel or company faced cases similar to yours.

Advanced out-of-the-box PTAB and TTAB analytics are always at your fingertips.

API

Docket Alarm offers a powerful API (application programming interface) to developers that want to integrate case filings into their apps.

LAW FIRMS
Build custom dashboards for your attorneys and clients with live data direct from the court.

Automate many repetitive legal tasks like conflict checks, document management, and marketing.

FINANCIAL INSTITUTIONS
Litigation and bankruptcy checks for companies and debtors.

E-DISCOVERY AND LEGAL VENDORS
Sync your system to PACER to automate legal marketing.