Review in Advance first posted online on April 1, 2013. (Changes may still occur before final publication online and in print.)

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 molecules.

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 *E. coli* Klenow fragment polymerase (more uniform incorporation of dideoxynucleotides) (10, 11);

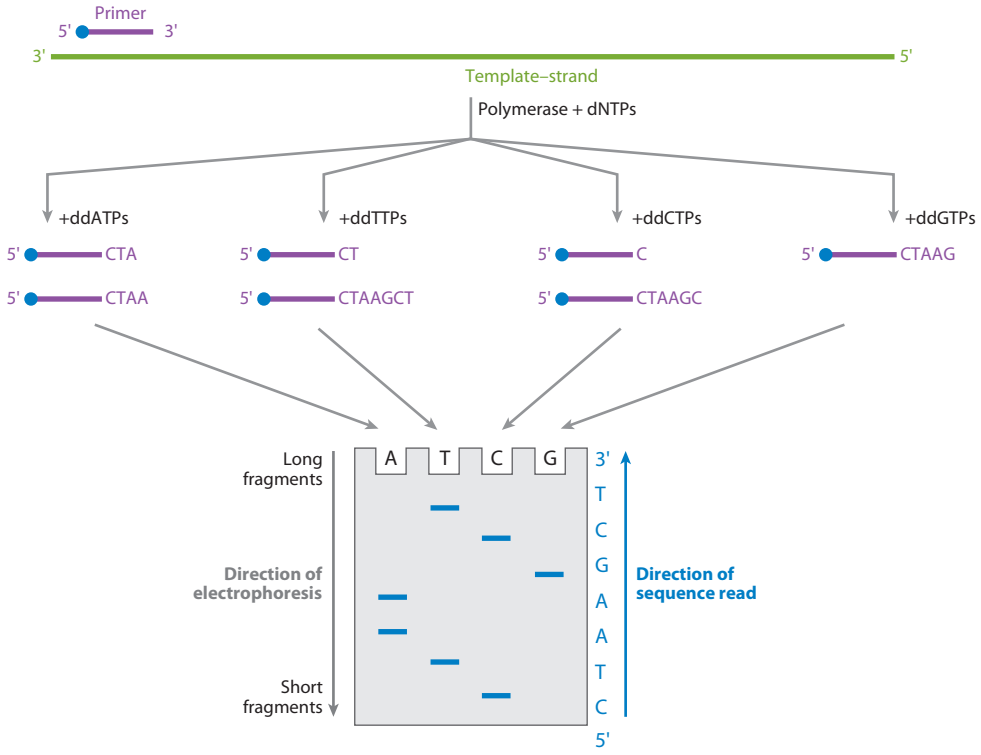

Figure 1

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, great increases were made in the daily throughput of sequencing data production, and several

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 clones by low-resolution mapping as a means to organize the high-resolution sequencing of smaller

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 entirely repetitive content score lowest due to the uncertainty of their origin (22, 23). Improved

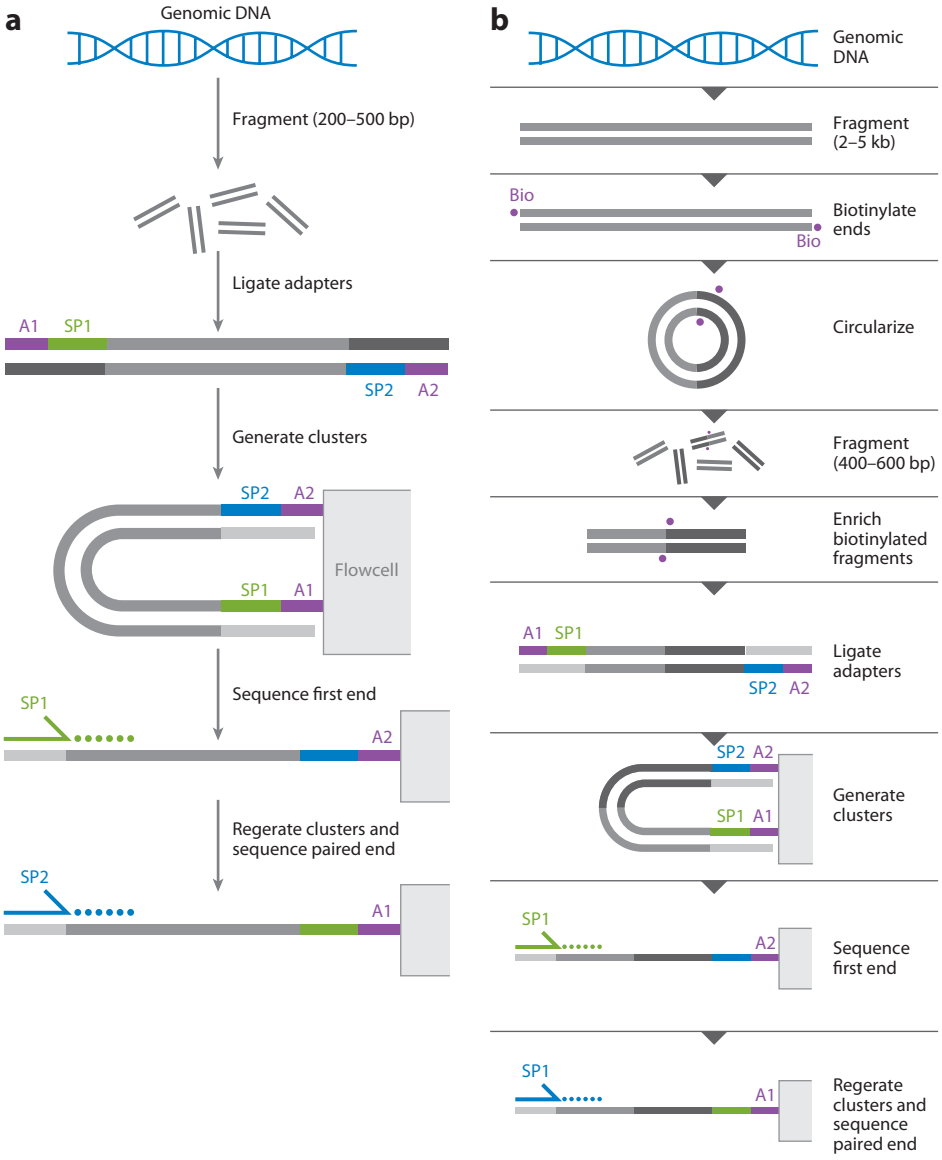

Figure 2

Comparison between (*a*) paired-end and (*b*) mate-pair sequencing library-construction processes.

certainty can be obtained from longer read lengths, and several next-generation sequencers have offered increases in read length over time and refinement of their signal-to-noise characteristics to allow this certainty. Another fundamental improvement has resulted from so-called paired-end sequencing, namely producing sequence data from both ends of each library fragment. Read pairs can be obtained by one of two mechanisms: (*a*) paired ends or (*b*) mate pairs (**Figure 2**).

In paired-end sequencing, a linear fragment with a length of less than 1 kb has adapter sequences at each end with different priming sites on each adapter. The sequencing instrument is designed to sequence from one adapter priming site by use of the stepwise sequencing described above; then,

in a subsequent reaction, the opposite adapter is primed and sequence data are obtained. These reads are paired with one another during the alignment step in data analysis, which provides higher overall certainty of placement than does a single end read of the same length. Most alignment algorithms also take into account the average length of fragments in the sequencing library to make the most accurate placement possible. In mate-pair sequencing, the library is constructed of fragments longer than 1 kb, and instead of ligating two adapters at each fragment end, the fragment is circularized around a single adapter and both fragment ends ligate to the adapter ends (24). These circular molecules are then treated by various molecular biology schemes (e.g., by type IIS endonuclease digestion or by nick translation) to produce a single linear fragment that holds both ends of the original DNA fragment with a central adapter. The remaining DNA remnants are removed by washing steps, as the central adapter that carries the mate-pair ends is biotinylated and can be captured using streptavidin magnetic beads. Typically, the resulting linear fragments have distinct adapters ligated to their ends, and sequencing is obtained from two sequential reads as described above. Again, the resulting reads are aligned as a pair to the genome of interest, wherein the separation distance between the reads is longer overall than that obtained with the paired-end approach. Often, mate-pair and paired-end reads are used in combination to achieve genome coverage when attempting longer-range assemblies through difficult regions of a genome or when attempting to assemble a genome for the first time (de novo sequencing) (25). In this combined coverage approach, the mate-pair reads provide longer-range order and orientation (a separation of up to 20 kb is possible), and the paired ends provide the ability to assemble, in a localized way, difficult-to-sequence regions that can then be layered on top of the scaffold provided by an assembly of mate-pair reads.

6. DIGITAL DATA TYPE AND RAMIFICATIONS

Next-generation sequencing libraries, carefully constructed to avoid sources of biasing and duplication, are highly digital. Specifically, the fact that each read originates from a consistently detected focus that results from the amplification of a single library fragment means that the data are inherently digital in nature. Thus, a quantitation of abundance can be inferred from this one-to-one relationship, which has ramifications for biological systems that are being investigated by next-generation sequencing. For example, chromosomal amplifications that are common in cancer genomes can be quantitated with respect to the extent of amplification (ploidy) on each chromosome (26). Similarly, the read prevalence of expressed genes identified by RNA sequencing can be directly correlated to their expression level and compared across replicates or with other samples from the same study (27). In population-based studies that use next-generation sequencing to characterize the individual species present in an isolate (metagenomics), a similar ability to correlate the presence of each species as a proportion of the overall population can be derived from the digital nature of next-generation sequencing data (28).

7. SOURCES OF NOISE AND ERROR MODELS

As mentioned above, although read length in next-generation sequencing is not limited by an electrophoretic separation step, the major limitation of read length is the signal-to-noise ratio during stepwise sequencing. Depending on the platform, the contributors to noise in the sequencing reaction differ, and there is interplay between the sources of noise and the sequencing errors that may result. This interplay gives rise to what is commonly referred to as the error model and is highly instrument and chemistry specific. In general, one typically explores both read-length limitations and error types by sequencing a reference set of genes or an entire genome, then comparing the

sequences obtained with the high-quality reference gene set or genome (29). In this approach, the different types of errors (substitution errors or insertion and deletion errors) can be identified, and the error model (random versus systematic errors) can be defined. Representation biases can also be uncovered by this approach when one examines the aligned reads for evidence of complete or partial lack of representation. If this lack of representation can be classified (for example, regions with >95% G + C content), then the bias can be defined. Typically, the more sequence reads are examined, the better defined are the error model, coverage biases, and their contributing sources. For example, the use of PCR or other types of enzymatic amplification may contribute systematic errors during the library construction or amplification processes described above. One might address this problem, independently of the instrument system used, by employing a high-fidelity polymerase and/or by limiting the number of amplification cycles when possible. Some sources of error, however, are simply instrument specific and may not be readily addressed by the end user (although they may improve over time with new chemistry and software from the manufacturer). As discussed below, instruments that use library amplification to enhance signal produced from the sequencing process forgo some of the signal-to-noise issues that are experienced in single-molecule systems because there are so many identical fragments being sequenced per focus that the number of fragments that are not misreporting far exceeds the number of fragments that are. In general, noise accumulates during the stepwise sequencing process and ultimately limits the read length obtained once the signal from any base incorporation step is outcompeted by incorrect or out-of-phase incorporation events, residual signal from prior reactions or reactants, and other sources of noise.

8. NEXT-GENERATION SEQUENCING WITH REVERSIBLE DYE TERMINATORS

It is informative to discuss some of the predominant approaches to next-generation sequencing as a means of tying together the concepts presented herein. The first instrument system involves the use of reversible dye terminators in enzymatic sequencing of amplified foci of library fragments. This system was initially developed in 2007 by Solexa and was subsequently acquired by Illumina®, Inc. (30). The library work flow follows steps similar to those outlined above, namely fragmentation of high–molecular weight DNA, enzymatic trimming, and adenylation of the fragment ends and ligation of specific adapters (**Figure 3*a***). The Illumina microfluidic conduit is a flow cell composed of flat glass with eight microfluidic channels, each decorated by covalent attachment of adapter sequences complementary to the library adapters. By careful quantitation of the library concentration, a precisely diluted solution of library fragments is amplified in situ on the flow cell surfaces by use of a bridge amplification step to produce foci for sequencing (clusters) (**Figure 3*b***). A subsequent step chemically effects the release of fragment ends carrying the same adapter, which is then primed with a complementary synthetic DNA (primer) to provide free 3′-OH groups that can be extended in subsequent stepwise sequencing reactions. In reversible dye terminator sequencing, all four nucleotides are provided in each cycle because each nucleotide carries an identifying fluorescent label. The sequencing occurs as single-nucleotide addition reactions because a blocking group exists at the 3′-OH position of the ribose sugar, preventing additional base incorporation reactions by the polymerase. As such, the series of events in each step includes the following, in order of occurrence: (*a*) The nucleotide is added by polymerase, (*b*) unincorporated nucleotides are washed away, (*c*) the flow cell is imaged on both inner surfaces to identify each cluster that is reporting a fluorescent signal, (*d*) the fluorescent groups are chemically cleaved, and (*e*) the 3′-OH is chemically deblocked (**Figure 3*c***). This series of steps is repeated for up to 150 nucleotide addition reactions, whereupon the second read preparations begin. To read from the opposite end of each fragment cluster, the instrument first removes the

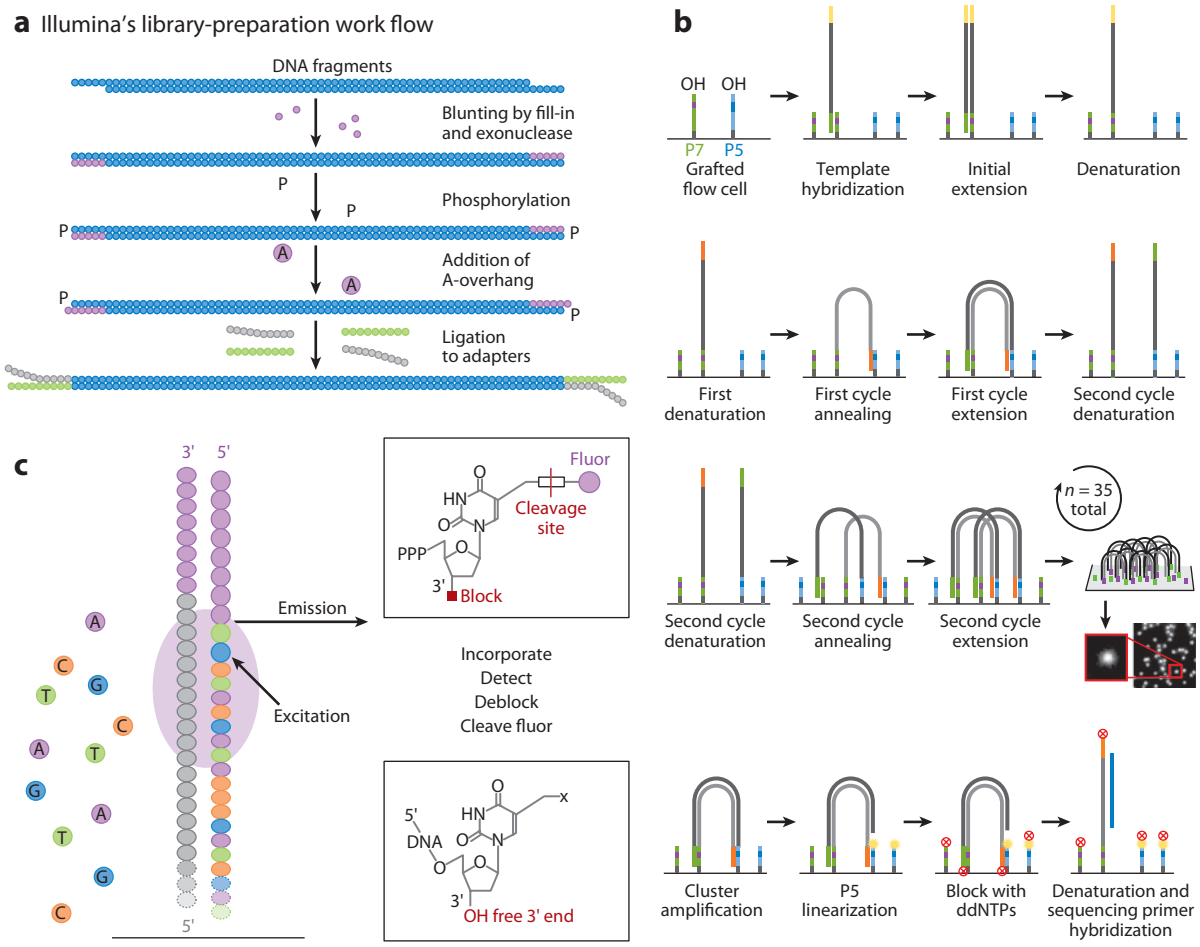

Figure 3

(*a*) Illumina® library-construction process. (*b*) Illumina cluster generation by bridge amplification. (*c*) Sequencing by synthesis with reversible dye terminators.

synthesized strands by denaturation and regenerates the clusters by performing a limited bridge amplification to improve the signal-to-noise ratio in the second read. After the amplification step, the opposite ends of the fragments are released from the flow cell surfaces by a different chemical cleavage reagent (corresponding to a labile group on the reverse adapter), and the fragments are primed with the reverse primer. Sequencing proceeds as described above. All of these steps occur on-instrument with the flow cell in place and without manual intervention, so the correlation of position from forward (first) to reverse (second) read is maintained and yields a very high read-pair concordance upon read alignment to the reference genome.

Illumina data have an error model that is described as having decreasing accuracy with increasing nucleotide addition steps. When errors occur, they are predominantly substitution errors, in which an incorrect nucleotide identity is assigned to the base. The error percentage of most Illumina reads is approximately 0.5% at best (i.e., 1 error in 200 bases). Sources of noise include (*a*) phasing, wherein increasing numbers of fragments fall out of phase with the majority

of fragments in the cluster due to incomplete deblocking in prior cycles or, conversely, due to lack of a blocking group that allows an additional base to be incorporated and (*b*) residual fluorescence interference noise due to incomplete fluorescent label cleavage from previous cycles.

Read lengths have increased from the original Solexa instrument at 25-bp single-end reads to the current Illumina HiSeq 2000 instrument's 150-bp paired-end reads. Increased read length has been one component that is contributing to an explosion in throughput-per-instrument run over a relatively short time frame (5 years), from 1 Gb for the Solexa 1G to 600 Gb for the HiSeq 2000. The latter instrument can thus produce sufficient data coverage for six whole–human genome sequences in approximately 11 days. The coverage per genome needed is approximately 30-fold, and with a 3-Gb genome wherein approximately 90% of the reads will map, 100 Gb are required to produce the necessary 90 Gb of data per genome.

The other contributor to throughput has been the ability to use increasingly more-concentrated library dilutions onto the flow cell, resulting in significant increases in cluster density. The HiSeq 2000 was the first instrument to read clusters from both surfaces of the flow cell channels, effectively doubling the throughput per run. Improvements in chemistry have made deblocking and fluor-removal steps more complete; polymerase engineering has improved incorporation fidelity and decreased errors and has decreased the G + C biases associated with the instrument at the bridge amplification step.

9. NEXT-GENERATION SEQUENCING BY pH CHANGE MONITORING

A completely different approach to next-generation sequencing is embodied in an instrument system that detects the release of hydrogen ions, a by-product of nucleotide incorporation, as quantitated changes in pH through a novel coupled silicon detector. This instrument was commercialized in 2010 by Ion Torrent (31), a company that was later purchased by Life Technologies™ Corp. For this approach, library construction includes DNA fragmentation, enzymatic end polishing, and adapter ligation. Amplification of library fragments occurs by a unique approach known as emulsion PCR, which quantitates the library fragments and dilutes them to be mixed in equimolar quantities with small beads, PCR reactants, and DNA polymerase molecules (32). The beads have covalently linked adapter complementary sequences on their surfaces to facilitate amplification on the bead. This mixture is then shaken to form an emulsion so that the beads and DNA are encapsulated in a 1:1 ratio (on average) in oil micelles that also contain the reactants needed for PCR-based amplification. The resulting mixture is placed into a specific apparatus that performs thermal cycling of the emulsion, effectively allowing hundreds of thousands of individual PCR amplifications to occur in parallel in one vessel. Subsequent steps are required first to separate the oil from the aqueous solution and beads (so-called emulsion breaking) and then to enrich the beads that were successfully amplified (to remove beads with insufficient DNA). Enriched beads are primed for sequencing by annealing a sequencing primer and are deposited into the wells of an Ion Chip, a specialized silicon chip designed to detect pH changes within individual wells of the sequencer as the reaction progresses stepwise. **Figure 4*a*** shows that the Ion Chip has an upper surface that serves as a microfluidic conduit to deliver the reactants needed for the sequencing reaction. The lower surface of the Ion Chip interfaces directly with a hydrogen ion detector that translates released hydrogen ions from each well into a quantitative readout of nucleotide bases that were incorporated in each reaction step (**Figure 4*b***). In this instrument, the reactant flow is by nucleotide in a systematic order because there is no label to provide base-specific identity upon incorporation. The adapter sequence contains a series of four single bases downstream of the primer's 3′-OH, in a sequence that matches the first four individual nucleotide flows across

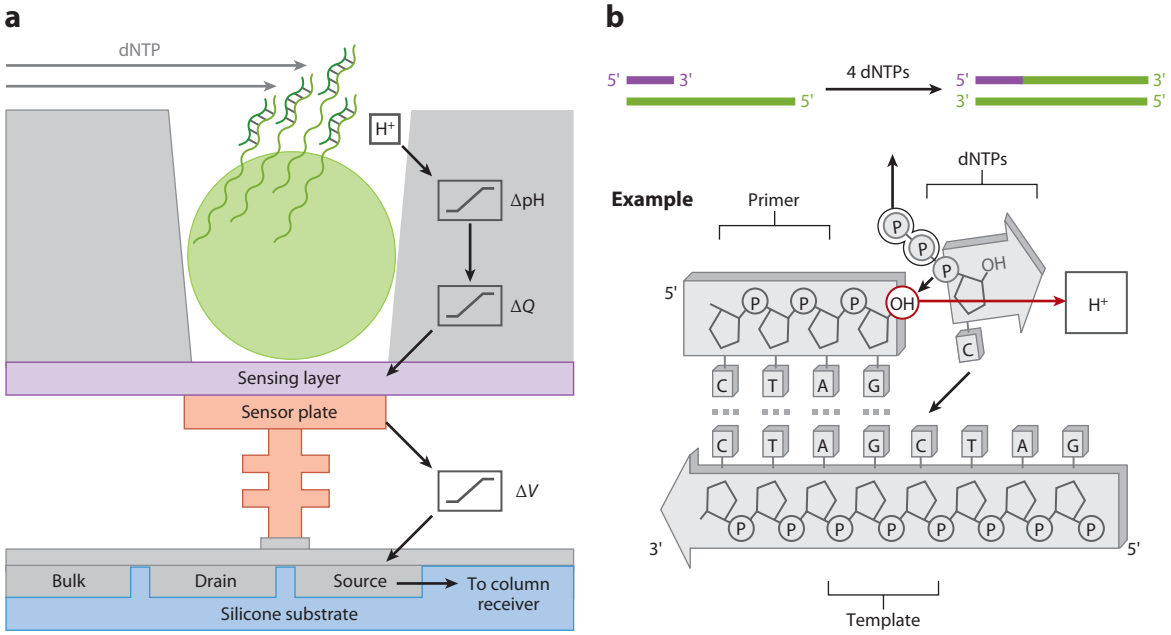

Figure 4

(*a*) Structure of the Ion Torrent Ion Chip used in pH-based sequencing. (*b*) pH sensing of nucleotide incorporation.

the chip, thereby providing a metric of single-base incorporation signal strength with which to calibrate the remaining responses during the ensuing incorporation steps.

Because the Ion Torrent sequencer uses native nucleotides for the sequencing reaction, there are no sources of noise akin to those identified for Illumina sequencing due to fluorescence or blocking groups on the reactants. Rather, noise accumulates due to phasing wherein not all the fragments are extended by nucleotide incorporation at each step. This effect is especially pronounced at sites in the library fragments with multiple bases of the same identity (so-called homopolymers). Coincidentally, the error model of Ion Torrent sequencing is defined largely by insertion or deletion errors that are also most prevalent at homopolymers. Here, the effect is most pronounced as the length of the homopolymer increases mainly due to loss of quantitation and ultimately saturation of the pH detector. Substitution errors also occur, albeit at very low frequency, and may be due to carryover effects from the previous incorporation cycle. Overall, the error rate of this instrument on a per read basis averages approximately 1% (i.e., 1 in 100 bases).

During the 2 years since the Ion Torrent was introduced, the average read length obtained has increased from 100 to 200 bp, produced as single-end reads. Unlike reversible terminator sequencing, the use of native nucleotides and the different sequences present on each bead loaded in the chip mean that incorporation rates differ from one bead to the next by incorporation cycle and according to sequence. As a result, a wide range of read lengths are obtained from any sequencing run, and this range increases as the total number of incorporation cycles increases. Throughput has increased over time, from 10 Mb per run average to 1 Gb per run average, by coupling longer reads with higher well density on the Ion Chip, which allows more beads to be sequenced per run. Each run requires approximately 2 h to complete; an intermediate series of washes requires an additional hour before the instrument can perform another run. Reaction volume miniaturization and the mass production of the Ion Chip using standard semiconductor techniques make this

technology relatively inexpensive and fast, and hence ideal for smaller laboratories that wish to use next-generation sequencing in their work but do not require extremely large data sets to accomplish it.

10. SINGLE-MOLECULE NEXT-GENERATION SEQUENCING: CHALLENGES

Although the advent of next-generation sequencing platforms has been a revolutionary advance in DNA sequencing throughput, there are difficulties associated with sequencing from polymerase-amplified fragment populations. These include the early occurrence of polymerase errors during library construction that may appear to be variant bases in the original genome but are not, as well as preferential amplification of certain fragments in the library population that cause them to be differentially overabundant relative to others. Furthermore, DNA modifications, such as different types of methylation, are diluted out during the amplification process because this step only copies the DNA bases. Therefore, advantages would exist for a platform that could obtain sequence data from individual molecules of a DNA isolate, given that many of the aforementioned challenges (e.g., polymerase errors and phasing) could be eliminated and effective read length increased as a result. However, there are unique challenges facing single-molecule DNA sequencing that, again, center largely around aspects of the signal-to-noise ratio but also relate directly to the fact that only one molecule per reaction is available for the production of signals from nucleotide incorporation. Therefore, the major challenge is to develop instrument detectors that can accurately interpret the extraordinarily low levels of signal produced from an individual molecule as it is being sequenced. Due to the inherently higher error rate of single-molecule sequencing for an individual read, mechanisms that permit multiple reads of the same molecule are inherently capable of producing an overall higher consensus read accuracy for the molecule.

11. SINGLE-MOLECULE NEXT-GENERATION SEQUENCING BY POLYMERASE ACTIVE-SITE MONITORING

One approach to single-molecule sequencing, commercialized in 2010 by Pacific Biosciences, Inc., combines nanotechnology with molecular biology and highly sensitive fluorescence detection to achieve single-molecule DNA sequencing (33). The nanotechnology employed in this platform is the zero-mode waveguide (ZMW). This light-focusing structure can be manufactured on a silicon wafer surface as a regular array that may contain tens to thousands of ZMWs. The molecular biology requires that DNA polymerase be significantly engineered to have the ideal properties for single-molecule sequencing; these properties include a decreased rate of polymerization and the ability to incorporate fluorescently modified nucleotides. The highly sensitive fluorescence detection is aimed at single-molecule detection of fluorescently labeled nucleotide incorporation events in real time, as the DNA polymerase is copying the template. By binding library fragments with DNA polymerase molecules, then depositing them by a diffusion-mediated process into ZMWs, the instrument excitation/detection optics can be trained on the bottom of each ZMW, where the polymerase attaches. Once fluorescent nucleotides are added to the chip surface, the instrument optics continually survey the active site of each ZMW-bound polymerase in real time, detecting the presence of each nucleotide that samples into the active site and dwells for a sufficient time to be incorporated into the growing DNA strand, by detecting the signal from its laser light–stimulated emission (**Figure 5**).

Although this platform monitors single molecules during sequencing, the library-preparation stages are very similar to those discussed above, with a few exceptions. DNA fragments are treated

Figure 5
Single-molecule sequencing using Pacific Biosciences' zero-mode waveguides.

to polish their ends, and adapters are ligated onto the ends. In the case of Pacific Biosciences, the adapters are hairpin adapters that, once ligated and denatured, form DNA circles consisting of adapter–Watson strand–adapter–Crick strands. After reaction by-products are removed, a primer that is complementary to the adapter is annealed to the library molecules, which are then complexed at a calculated ratio with DNA polymerase molecules. This mixture is supplied to the instrument, which deposits a portion of the mixture onto the surface of the ZMW chip (SMRT® Cell) that has been carefully placed onto the optical stage, such that each excitation laser beamlet is aligned with each of the 150,000 ZMWs on the SMRT Cell, and then adds the sequencing reactants. Subsequently, the instrument performs the sequencing data collection by actively monitoring each ZMW to record incorporating fluorescent molecules dwelling in the active site as identified by their emission wavelength. Because the nucleotides used by this platform carry their fluorescent label on the pyrophosphate, the fluorescent group diffuses away once released in the incorporation reaction.

Due to the large number of ZMWs being monitored constantly, the instrument computer cluster performs numerous calculations during the run to effectively condense the data into a workable file size that can be subsequently analyzed to produce the sequencing reads. Depending on the size of the library fragments and the time of data collection, the read lengths obtained can be quite long (up to 10,000 nucleotides at maximum) or can provide multiple passes through the Watson and Crick strands of shorter inserts. There are, as mentioned above, unique factors that contribute to an overall higher error rate in single-molecule sequencing, such as that provided by Pacific Biosciences. For example, some small fraction of nucleotides escape fluorescent labeling and subsequent purification steps and do not register a base incorporation when they are added into the synthesized strand. Another source of error originates in nucleotides that dwell long enough to be detected but are not incorporated and, conversely, in nucleotides that do not dwell long enough to be detected yet are incorporated into the synthesized strand. The number of nucleotides of a given identity may also be incorrect when the nucleotide dwell time is longer than average, such that multiple incorporations appear to be occurring but are not. Cumulatively, these errors dictate that the error model for individual reads is predominated by insertion and deletion errors, although some smaller proportion of substitution errors do occur. On a per read basis, the error rate is approximately 15% (15 bases per 100), yet the consensus error rate obtained from multiple passes on a single molecule is >97%. Recent research on specialized analysis of the incorporation rate of specifically methylated DNA has indicated that the Pacific Biosciences platform can perform sequencing and coincident detection of specific methyl residues known to reside on DNA as a regulatory mechanism (34, 35). Other investigators have used the Pacific Biosciences platform to measure, at single-molecule resolution, the mechanism involved in ribosomal translocation on a messenger RNA molecule (36) and to examine the structures of individual protein receptors on the plasma membranes of living cells (37).

12. CONCLUSIONS

The advent of next-generation sequencing has fueled a revolution in biological research that has largely capitalized on the previous decade of research, which established genome reference sequences for many model organisms and the human. These advances arise mainly from the significant lowering of cost and the remarkable increase in data-production capacity, as described in this review. In effect, the sequencing of a human genome can now be accomplished in the data-generation phase within 2 weeks at a cost of approximately $5,000 (on average). As this timeline for data production decreases to a single day, and as the cost continues to fall, the use of next-generation sequencing technology in the clinical setting is becoming a pressing issue.

Beyond human genetic diseases, however, the impact of next-generation sequencing has been similarly profound. For example, several "big science" projects have benefited from the low cost and high throughput of next-generation sequencing. One such project, the ENCODE project, recently reported the results of genome-wide annotation of functional sites in mouse and human, effectively increasing our knowledge of the regulatory sequences within these genomes (38–41). Another example is the Human Microbiome Project, which has used next-generation sequencing to characterize the diversity and types of bacteria and viruses that inhabit various areas of the human body in several thousand healthy individuals (28, 42, 43). These projects have established a well-defined baseline for microbial health, which can subsequently inform the changes to microbial population shifts in the context of disease.

Next-generation sequencing has also strongly influenced the experimental cost and scope for individual laboratories. The sequencing of phenotyped strains can be rapidly characterized at the genomic level to, for example, identify putative genes whose mutational status contribute to the phenotype, evaluate changes in genome-wide methylation patterns, or readily examine changes in gene expression that are a consequence of a single genome-level mutation. Taken together, the continuing trends in data-generation facility and cost reduction in next-generation sequencing platforms will probably contribute, over the long term, to increasing our genome-wide knowledge of organisms and organism systems.

DISCLOSURE STATEMENT

The author is a member of the Scientific Advisory Board for Pacific Biosciences, Inc.; a speakers bureau member for Illumina, Inc.; and a shareholder in Life Technologies Corp.

LITERATURE CITED

1. Sanger F, Coulson AR. 1975. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J. Mol. Biol.* 94:441–48
2. Sanger F, Donelson JE, Coulson AR, Kossel H, Fischer D. 1974. Determination of a nucleotide sequence in bacteriophage f1 DNA by primed synthesis with DNA polymerase. *J. Mol. Biol.* 90:315–33
3. Sanger F, Donelson JE, Coulson AR, Kossel H, Fischer D. 1973. Use of DNA polymerase I primed by a synthetic oligonucleotide to determine a nucleotide sequence in phage fl DNA. *Proc. Natl. Acad. Sci. USA* 70:1209–13
4. Barrell BG, Sanger F. 1969. The sequence of phenylalanine tRNA from *E. coli*. *FEBS Lett.* 3:275–78
5. Sanger F, Brownlee GG, Barrell BG. 1965. A two-dimensional fractionation procedure for radioactive nucleotides. *J. Mol. Biol.* 13:373–98
6. Brownlee GG, Sanger F, Barrell BG. 1967. Nucleotide sequence of 5S-ribosomal RNA from *Escherichia coli*. *Nature* 215:735–36
7. Sanger F, Nicklen S, Coulson AR. 1977. DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. USA* 74:5463–67
8. Gait MJ, Matthes HW, Singh M, Sproat BS, Titmas RC. 1982. Rapid synthesis of oligodeoxyribonucleotides. VII. Solid phase synthesis of oligodeoxyribonucleotides by a continuous flow phosphotriester method on a kieselguhr–polyamide support. *Nucleic Acids Res.* 10:6243–54
9. Gait MJ, Sheppard RC. 1977. Rapid synthesis of oligodeoxyribonucleotides: a new solid-phase method. *Nucleic Acids Res.* 4:1135–58
10. Tabor S, Huber HE, Richardson CC. 1987. *Escherichia coli* thioredoxin confers processivity on the DNA polymerase activity of the gene 5 protein of bacteriophage T7. *J. Biol. Chem.* 262:16212–23
11. Tabor S, Richardson CC. 1995. A single residue in DNA polymerases of the *Escherichia coli* DNA polymerase I family is critical for distinguishing between deoxy- and dideoxyribonucleotides. *Proc. Natl. Acad. Sci. USA* 92:6339–43

12. Smith LM, Sanders JZ, Kaiser RJ, Hughes P, Dodd C, et al. 1986. Fluorescence detection in automated DNA sequence analysis. *Nature* 321:674–79
13. Fulton LL, Wilson RK. 1994. Variations on cycle sequencing. *Biotechniques* 17:298–301
14. McBride LJ, Koepf SM, Gibbs RA, Salser W, Mayrand PE, et al. 1989. Automated DNA sequencing methods involving polymerase chain reaction. *Clin. Chem.* 35:2196–201
15. Saiki RK, Scharf S, Faloona F, Mullis KB, Horn GT, et al. 1985. Enzymatic amplification of β-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. *Science* 230:1350–54
16. Prober JM, Trainor GL, Dam RJ, Hobbs FW, Robertson CW, et al. 1987. A system for rapid DNA sequencing with fluorescent chain–terminating dideoxynucleotides. *Science* 238:336–41
17. Panussis DA, Cook MW, Rifkin LL, Snider JE, Strong JT, et al. 1998. A pneumatic device for rapid loading of DNA sequencing gels. *Genome Res.* 8:543–48
18. Marsh M, Tu O, Dolnik V, Roach D, Solomon N, et al. 1997. High-throughput DNA sequencing on a capillary array electrophoresis system. *J. Capill. Electrophor.* 4:83–89
19. Mardis ER. 2011. A decade's perspective on DNA sequencing technology. *Nature* 470:198–203
20. Metzker ML. 2010. Sequencing technologies—the next generation. *Nat. Rev. Genet.* 11:31–46
21. Hsiao KK, Cass C, Schellenberg GD, Bird T, Devine-Gage E, et al. 1991. A prion protein variant in a family with the telencephalic form of Gerstmann–Straussler–Scheinker syndrome. *Neurology* 41:681–84
22. Li H, Durbin R. 2010. Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics* 26:589–95
23. Li H, Ruan J, Durbin R. 2008. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.* 18:1851–58
24. Korbel JO, Urban AE, Affourtit JP, Godwin B, Grubert F, et al. 2007. Paired-end mapping reveals extensive structural variation in the human genome. *Science* 318:420–26
25. Gnerre S, Maccallum I, Przybylski D, Ribeiro FJ, Burton JN, et al. 2011. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proc. Natl. Acad. Sci. USA* 108:1513–18
26. Zhang Q, Ding L, Larson DE, Koboldt DC, McLellan MD, et al. 2010. CMDS: a population-based method for identifying recurrent DNA copy number aberrations in cancer from high-resolution data. *Bioinformatics* 26:464–69
27. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* 5:621–28
28. Hum. Microbiome Proj. Consort. 2012. Structure, function and diversity of the healthy human microbiome. *Nature* 486:207–14
29. Hillier LW, Marth GT, Quinlan AR, Dooling D, Fewell G, et al. 2008. Whole-genome sequencing and variant discovery in *C. elegans*. *Nat. Methods* 5:183–88
30. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, et al. 2008. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* 456:53–59
31. Rothberg JM, Hinz W, Rearick TM, Schultz J, Mileski W, et al. 2011. An integrated semiconductor device enabling non-optical genome sequencing. *Nature* 475:348–52
32. Dressman D, Yan H, Traverso G, Kinzler KW, Vogelstein B. 2003. Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations. *Proc. Natl. Acad. Sci. USA* 100:8817–22
33. Eid J, Fehr A, Gray J, Luong K, Lyle J, et al. 2009. Real-time DNA sequencing from single polymerase molecules. *Science* 323:133–38
34. Song CX, Clark TA, Lu XY, Kislyuk A, Dai Q, et al. 2012. Sensitive and specific single-molecule sequencing of 5-hydroxymethylcytosine. *Nat. Methods* 9:75–77
35. Murray IA, Clark TA, Morgan RD, Boitano M, Anton BP, et al. 2012. The methylomes of six bacteria. *Nucleic Acids Res.* 40:11450–62
36. Uemura S, Aitken CE, Korlach J, Flusberg BA, Turner SW, Puglisi JD. 2010. Real-time tRNA transit on single translating ribosomes at codon resolution. *Nature* 464:1012–17
37. Richards CI, Luong K, Srinivasan R, Turner SW, Dougherty DA, et al. 2012. Live-cell imaging of single receptor composition using zero-mode waveguide nanostructures. *Nano Lett.* 12:3690–94
38. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, et al. 2012. GENCODE: the reference human genome annotation for the ENCODE Project. *Genome Res.* 22:1760–74

39. Bernstein BE, Birney E, Dunham I, Green ED, Gunter C, Snyder M. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* 489:57–74
40. Yip KY, Cheng C, Bhardwaj N, Brown JB, Leng J, et al. 2012. Classification of human genomic regions based on experimentally determined binding sites of more than 100 transcription-related factors. *Genome Biol.* 13:R48
41. Stamatoyannopoulos JA, Snyder M, Hardison R, Ren B, Gingeras T, et al. 2012. An encyclopedia of mouse DNA elements (Mouse ENCODE). *Genome Biol.* 13:418
42. Gevers D, Knight R, Petrosino JF, Huang K, McGuire AL, et al. 2012. The human microbiome project: a community resource for the healthy human microbiome. *PLoS Biol.* 10:e1001377
43. Abubucker S, Segata N, Goll J, Schubert AM, Izard J, et al. 2012. Metabolic reconstruction for metagenomic data and its application to the human microbiome. *PLoS Comput. Biol.* 8:e1002358