APPLICATIONS OF NEXT-GENERATION SEQUENCING

EXHIBIT 2053
FLOYD ROMESBERG, PH.D.
09-19-18
Patricia Y. Schuler, CSR No. 11949
APTUS COURT REPORTING

Sequencing technologies — the next generation

*Michael L. Metzker**‡*

Abstract | Demand has never been greater for revolutionary technologies that deliver fast, inexpensive and accurate genome information. This challenge has catalysed the development of next-generation sequencing (NGS) technologies. The inexpensive production of large volumes of sequence data is the primary advantage over conventional methods. Here, I present a technical review of template preparation, sequencing and imaging, genome alignment and assembly approaches, and recent advances in current and near-term commercially available NGS instruments. I also outline the broad range of applications for NGS technologies, in addition to providing guidelines for platform selection to address biological questions of interest.

Automated Sanger sequencing
This process involves a mixture of techniques: bacterial cloning or PCR; template purification; labelling of DNA fragments using the chain termination method with energy transfer, dye-labelled dideoxynucleotides and a DNA polymerase; capillary electrophoresis; and fluorescence detection that provides four-colour plots to reveal the DNA sequence.

Over the past four years, there has been a fundamental shift away from the application of automated Sanger sequencing for genome analysis. Prior to this departure, the automated Sanger method had dominated the industry for almost two decades and led to a number of monumental accomplishments, including the completion of the only finished-grade human genome sequence[1]. Despite many technical improvements during this era, the limitations of automated Sanger sequencing showed a need for new and improved technologies for sequencing large numbers of human genomes. Recent efforts have been directed towards the development of new methods, leaving Sanger sequencing with fewer reported advances. As such, automated Sanger sequencing is not covered here, and interested readers are directed to previous articles[2,3].

The automated Sanger method is considered as a 'first-generation' technology, and newer methods are referred to as next-generation sequencing (NGS). These newer technologies constitute various strategies that rely on a combination of template preparation, sequencing and imaging, and genome alignment and assembly methods. The arrival of NGS technologies in the marketplace has changed the way we think about scientific approaches in basic, applied and clinical research. In some respects, the potential of NGS is akin to the early days of PCR, with one's imagination being the primary limitation to its use. The major advance offered by NGS is the ability to produce an enorm volume of data cheaply — in some cases in excess one billion short reads per instrument run. This feat expands the realm of experimentation beyond j determining the order of bases. For example, in gene-expression studies microarrays are now being replaced by seq-based methods, which can identify and quantify rare transcripts without prior knowledge of a particular gene and can provide information regarding alternative splicing and sequence variation in identified genes[4,5]. The ability to sequence the whole genome of many related organisms has allowed large-scale comparative and evolutionary studies to be performed that were unimaginable just a few years ago. The broadest application of NGS may be the resequencing of human genomes to enhance our understanding of how genetic differences affect health and disease. The variety of NGS features makes it likely that multiple platforms will coexist in the marketplace, with some having clear advantages for particular applications over others.

This Review focuses on commercially available technologies from Roche/454, Illumina/Solexa, Life/APG and Helicos BioSciences, the Polonator instrument and the near-term technology of Pacific Biosciences, who aim to bring their sequencing device to the market in 2010. Nanopore sequencing is not covered, although interested readers are directed to an article by Branton and colleagues[6], who describe the advances and remaining challenges for this technology. Here, I present a technical review of template preparation, sequencing and imaging, genome alignment and assembly, and current NGS platform performance to provide guidance on how

*Human Genome Sequencing Center and Department of Molecular & Human Genetics, Baylor College of Medicine, One Baylor Plaza, N1409, Houston, Texas 77030, USA.
‡LaserGen, Inc., 8052 El Rio Street, Houston, Texas 77054, USA.
e-mail: mmetzker@bcm.edu
doi:10.1038/nrg2626
Published online
8 December 2009

©2010 Macmillan Publishers Limited.

Columbia Exhibit 2053
Illumina, Inc. v. The Trustees of Columbia University in the City of New York
IPR2018-00291, 318, 322 & 385

APPLICATIONS OF NEXT-GENERATION SEQUENCING

Sequencing technologies — the next generation

*Michael L. Metzker**‡

Abstract | Demand has never been greater for revolutionary technologies that deliver fast, inexpensive and accurate genome information. This challenge has catalysed the development of next-generation sequencing (NGS) technologies. The inexpensive production of large volumes of sequence data is the primary advantage over conventional methods. Here, I present a technical review of template preparation, sequencing and imaging, genome alignment and assembly approaches, and recent advances in current and near-term commercially available NGS instruments. I also outline the broad range of applications for NGS technologies, in addition to providing guidelines for platform selection to address biological questions of interest.

Automated Sanger sequencing
This process involves a mixture of techniques: bacterial cloning or PCR; template purification; labelling of DNA fragments using the chain termination method with energy transfer, dye-labelled dideoxynucleotides and a DNA polymerase; capillary electrophoresis; and fluorescence detection that provides four-colour plots to reveal the DNA sequence.

*Human Genome Sequencing Center and Department of Molecular & Human Genetics, Baylor College of Medicine, One Baylor Plaza, N1409, Houston, Texas 77030, USA.
‡LaserGen, Inc., 8052 El Rio Street, Houston, Texas 77054, USA.
e-mail: mmetzker@bcm.edu
doi:10.1038/nrg2626
Published online
8 December 2009

Over the past four years, there has been a fundamental shift away from the application of automated Sanger sequencing for genome analysis. Prior to this departure, the automated Sanger method had dominated the industry for almost two decades and led to a number of monumental accomplishments, including the completion of the only finished-grade human genome sequence[1]. Despite many technical improvements during this era, the limitations of automated Sanger sequencing showed a need for new and improved technologies for sequencing large numbers of human genomes. Recent efforts have been directed towards the development of new methods, leaving Sanger sequencing with fewer reported advances. As such, automated Sanger sequencing is not covered here, and interested readers are directed to previous articles[2,3].

The automated Sanger method is considered as a 'first-generation' technology, and newer methods are referred to as next-generation sequencing (NGS). These newer technologies constitute various strategies that rely on a combination of template preparation, sequencing and imaging, and genome alignment and assembly methods. The arrival of NGS technologies in the marketplace has changed the way we think about scientific approaches in basic, applied and clinical research. In some respects, the potential of NGS is akin to the early days of PCR, with one's imagination being the primary limitation to its use. The major advance offered by NGS is the ability to produce an enormous volume of data cheaply — in some cases in excess of one billion short reads per instrument run. This feature expands the realm of experimentation beyond just determining the order of bases. For example, in gene-expression studies microarrays are now being replaced by seq-based methods, which can identify and quantify rare transcripts without prior knowledge of a particular gene and can provide information regarding alternative splicing and sequence variation in identified genes[4,5]. The ability to sequence the whole genome of many related organisms has allowed large-scale comparative and evolutionary studies to be performed that were unimaginable just a few years ago. The broadest application of NGS may be the resequencing of human genomes to enhance our understanding of how genetic differences affect health and disease. The variety of NGS features makes it likely that multiple platforms will coexist in the marketplace, with some having clear advantages for particular applications over others.

This Review focuses on commercially available technologies from Roche/454, Illumina/Solexa, Life/APG and Helicos BioSciences, the Polonator instrument and the near-term technology of Pacific Biosciences, who aim to bring their sequencing device to the market in 2010. Nanopore sequencing is not covered, although interested readers are directed to an article by Branton and colleagues[6], who describe the advances and remaining challenges for this technology. Here, I present a technical review of template preparation, sequencing and imaging, genome alignment and assembly, and current NGS platform performance to provide guidance on how these technologies work and how they may be applied to important biological questions. I highlight the applications of human genome resequencing using targeted and whole-genome approaches, and discuss the progress

©2010 Macmillan Publishers Limited. All rights reserved

and limitations of these methods, as well as upcoming advances and the impact they are expected to have over the next few years.

Next-generation sequencing technologies

Sequencing technologies include a number of methods that are grouped broadly as template preparation, sequencing and imaging, and data analysis. The unique combination of specific protocols distinguishes one technology from another and determines the type of data produced from each platform. These differences in data output present challenges when comparing platforms based on data quality and cost. Although quality scores and accuracy estimates are provided by each manufacturer, there is no consensus that a 'quality base' from one platform is equivalent to that from another platform. Various sequencing metrics are discussed later in the article.

In the following sections, stages of template preparation and sequencing and imaging are discussed as they apply to existing and near-term commercial platforms. There are two methods used in preparing templates for NGS reactions: clonally amplified templates originating from single DNA molecules, and single DNA-molecule templates. The term sequencing by synthesis, which is used to describe numerous DNA polymerase-dependent methods in the literature, is not used in this article because it fails to delineate the different mechanisms involved in sequencing[2,7]. Instead, these methods are classified as cyclic reversible termination (CRT), single-nucleotide addition (SNA) and real-time sequencing. Sequencing by ligation (SBL), an approach in which DNA polymerase is replaced by DNA ligase, is also described. Imaging methods coupled with these sequencing strategies range from measuring bioluminescent signals to four-colour imaging of single molecular events. The voluminous data produced by these NGS platforms place substantial demands on information technology in terms of data storage, tracking and quality control (see REF. 8 for details).

Template preparation

The need for robust methods that produce a representative, non-biased source of nucleic acid material from the genome under investigation cannot be overemphasized. Current methods generally involve randomly breaking genomic DNA into smaller sizes from which either fragment templates or mate-pair templates are created. A common theme among NGS technologies is that the template is attached or immobilized to a solid surface or support. The immobilization of spatially separated template sites allows thousands to billions of sequencing reactions to be performed simultaneously.

Clonally amplified templates. Most imaging systems have not been designed to detect single fluorescent events, so amplified templates are required. The two most common methods are emulsion PCR (emPCR)[9] and solid-phase amplification[10]. EmPCR is used to prepare sequencing templates in a cell-free system, which has the advantage of avoiding the arbitrary loss of genomic sequences — a problem that is inherent in bacterial cloning methods. A library of fragment or mate-pair targets is created, and adaptors containing universal priming sites are ligated to the target ends, allowing complex genomes to be amplified with common PCR primers. After ligation, the DNA is separated into single strands and captured onto beads under conditions that favour one DNA molecule per bead (FIG. 1a). After the successful amplification and enrichment of emPCR beads, millions can be immobilized in a polyacrylamide gel on a standard microscope slide (Polonator)[11], chemically crosslinked to an amino-coated glass surface (Life/APG; Polonator)[12] or deposited into individual PicoTiterPlate (PTP) wells (Roche/454)[13] in which the NGS chemistry can be performed.

Solid-phase amplification can also be used to produce randomly distributed, clonally amplified clusters from fragment or mate-pair templates on a glass slide (FIG. 1b). High-density forward and reverse primers are covalently attached to the slide, and the ratio of the primers to the template on the support defines the surface density of the amplified clusters. Solid-phase amplification can produce 100–200 million spatially separated template clusters (Illumina/Solexa), providing free ends to which a universal sequencing primer can be hybridized to initiate the NGS reaction.

Single-molecule templates. Although clonally amplified methods offer certain advantages over bacterial cloning, some of the protocols are cumbersome to implement and require a large amount of genomic DNA material (3–20 μg). The preparation of single-molecule templates is more straightforward and requires less starting material (<1 μg). More importantly, these methods do not require PCR, which creates mutations in clonally amplified templates that masquerade as sequence variants. AT-rich and GC-rich target sequences may also show amplification bias in product yield, which results in their underrepresentation in genome alignments and assemblies. Quantitative applications, such as RNA–seq[5], perform more effectively with non-amplified template sources, which do not alter the representational abundance of mRNA molecules.

Before the NGS reaction is carried out, single-molecule templates are usually immobilized on solid supports using one of at least three different approaches. In the first approach, spatially distributed individual primer molecules are covalently attached to the solid support[14]. The template, which is prepared by randomly fragmenting the starting material into small sizes (for example, ~200–250 bp) and adding common adaptors to the fragment ends, is then hybridized to the immobilized primer (FIG. 1c). In the second approach, spatially distributed single-molecule templates are covalently attached to the solid support[14] by priming and extending single-stranded, single-molecule templates from immobilized primers (FIG. 1c). A common primer is then hybridized to the template (FIG. 1d). In either approach, DNA polymerase can bind to the immobilized primed template configuration to initiate the NGS reaction. Both of the above approaches are used by Helicos BioSciences. In a third approach, spatially distributed single polymerase molecules

Finished grade
A quality measure for a sequenced genome. A finished-grade genome, commonly referred to as a 'finished genome', is of higher quality than a draft-grade genome, with more base coverage and fewer errors and gaps (for example, the human genome reference contains 2.85 Gb, covers 99% of the genome with 341 gaps, and has an error rate of 1 in every 100,000 bp).

Template
This recombinant DNA molecule is made up of a known region, usually a vector or adaptor sequence to which a universal primer can bind, and the target sequence, which is typically an unknown portion to be sequenced.

Seq-based methods
Assays that use next-generation sequencing technologies. They include methods for determining the sequence content and abundance of mRNAs, non-coding RNAs and small RNAs (collectively called RNA–seq) and methods for measuring genome-wide profiles of immunoprecipitated DNA–protein complexes (ChIP–seq), methylation sites (methyl–seq) and DNase I hypersensitivity sites (DNase–seq).

Polonator
This Review mostly describes technology platforms that are associated with a respective company, but the Polonator G.007 instrument, which is manufactured and distributed by Danaher Motions (a Dover Company), is an open source platform with freely available software and protocols. Users manufacture their own reagents based on published reports or by collaborating with George Church and colleagues or other technology developers.

Fragment templates
A fragment library is prepared by randomly shearing genomic DNA into small sizes of <1 kb, and requires less DNA than would be needed for a mate-pair library.

©2010 Macmillan Publishers Limited. All rights reserved

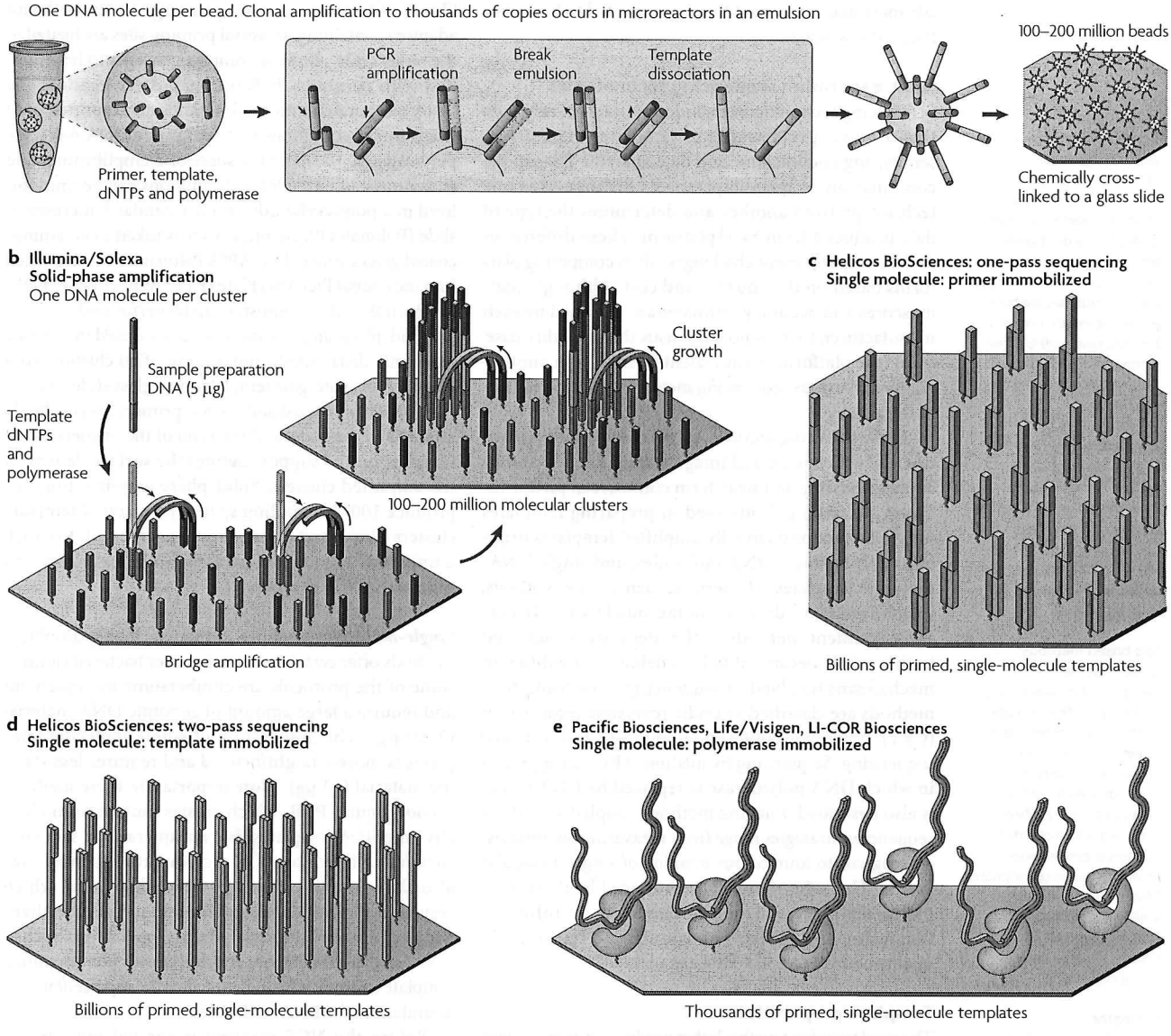


Figure 1 | **Template immobilization strategies.** In emulsion PCR (emPCR) (**a**), a reaction mixture consisting of an oil–aqueous emulsion is created to encapsulate bead–DNA complexes into single aqueous droplets. PCR amplification is performed within these droplets to create beads containing several thousand copies of the same template sequence. EmPCR beads can be chemically attached to a glass slide or deposited into PicoTiterPlate wells (FIG. 3c). Solid-phase amplification (**b**) is composed of two basic steps: initial priming and extending of the single-stranded, single-molecule template, and bridge amplification of the immobilized template with immediately adjacent primers to form clusters. Three approaches are shown for immobilizing single-molecule templates to a solid support: immobilization by a primer (**c**); immobilization by a template (**d**); and immobilization of a polymerase (**e**). dNTP, 2′-deoxyribonucleoside triphosphate.

Mate-pair templates
A genomic library is prepared by circularizing sheared DNA that has been selected for a given size, such as 2 kb, therefore bringing the ends that were previously distant from one another into close proximity. Cutting these circles into linear DNA fragments creates mate-pair templates.

are attached to the solid support[15], to which a primed template molecule is bound (FIG. 1e). This approach is used by Pacific Biosciences[15] and is described in patents from Life/VisiGen[16] and LI-COR Biosciences[17]. Larger DNA molecules (up to tens of thousands of base pairs) can be used with this technique and, unlike the first two approaches, the third approach can be used with real-time methods, resulting in potentially longer read lengths.

Sequencing and imaging

There are fundamental differences in sequencing clonally amplified and single-molecule templates. Clonal amplification results in a population of identical templates, each of which has undergone the sequencing reaction. Upon imaging, the observed signal is a consensus of the nucleotides or probes added to the identical templates for a given cycle. This places a greater

©2010 Macmillan Publishers Limited. All rights reserved

demand on the efficiency of the addition process, and incomplete extension of the template ensemble results in lagging-strand dephasing. The addition of multiple nucleotides or probes can also occur in a given cycle, resulting in leading-strand dephasing. Signal dephasing increases fluorescence noise, causing base-calling errors and shorter reads[18]. Because dephasing is not an issue with single-molecule templates, the requirement for cycle efficiency is relaxed. Single molecules, however, are susceptible to multiple nucleotide or probe additions in any given cycle. Here, deletion errors will occur owing to quenching effects between adjacent dye molecules or no signal will be detected because of the incorporation of dark nucleotides or probes. In the following sections, sequencing and imaging strategies that use both clonally amplified and single-molecule templates are discussed.

Cyclic reversible termination. As the name implies, CRT uses reversible terminators in a cyclic method that comprises nucleotide incorporation, fluorescence imaging and cleavage[2]. In the first step, a DNA polymerase, bound to the primed template, adds or incorporates just one fluorescently modified nucleotide (BOX 1), which represents the complement of the template base. The termination of DNA synthesis after the addition of a single nucleotide is an important feature of CRT. Following incorporation, the remaining unincorporated nucleotides are washed away. Imaging is then performed to determine the identity of the incorporated nucleotide. This is followed by a cleavage step, which removes the terminating/inhibiting group and the fluorescent dye. Additional washing is performed before starting the next incorporation step. FIG. 2a depicts a four-colour CRT cycle used by Illumina/Solexa, and FIG. 2c illustrates a one-colour CRT cycle used by Helicos BioSciences.

The key to the CRT method is the reversible terminator, of which there are two types: 3′ blocked and 3′ unblocked (BOX 1). The use of a dideoxynucleotide, which acts as a chain terminator in Sanger sequencing, provided the basis for the initial development of reversible blocking groups attached to the 3′ end of nucleotides[19,20]. Blocking groups, such as 3′-*O*-allyl-2′-deoxyribonucleoside triphosphates (dNTPs)[21] and 3′-*O*-azidomethyl-dNTPs[22], have been successfully used in CRT. 3′-blocked terminators require the cleavage of two chemical bonds to remove the fluorophore from the nucleobase and restore the 3′-OH group.

Currently, the Illumina/Solexa Genome Analyzer (GA)[23] dominates the NGS market. It uses the clonally amplified template method illustrated in FIG. 1b, coupled with the four-colour CRT method illustrated in FIG. 2a. The four colours are detected by total internal reflection fluorescence (TIRF) imaging using two lasers, the output of which is depicted in FIG. 2b. The slide is partitioned into eight channels, which allows independent samples to be run simultaneously. TABLE 1 shows the current sequencing statistics of the Illumina/Solexa GA_{II} platform operating at the Baylor College of Medicine Human Genome Sequencing Center (BCM-HGSC; D. Muzny, personal communication). Substitutions are the most common error type, with a higher portion of errors occurring when the previous incorporated nucleotide is a 'G' base[24]. Genome analysis of Illumina/Solexa data has revealed an underrepresentation of AT-rich[24–26] and GC-rich regions[25,26], which is probably due to amplification bias during template preparation[25]. Sequence variants are called by aligning reads to a reference genome using bioinformatics tools such as MAQ[27] or ELAND[23]. Bentley and colleagues reported high concordance (>99.5%) of single-nucleotide variant (SNV)[28] calls with standard genotyping arrays using both alignment tools, and a false-positive rate of 2.5% with novel SNVs[23]. Other reports have described a higher false-positive rate associated with novel SNV detection using these alignment tools[29,30].

The difficulty involved in identifying a modified enzyme that efficiently incorporates 3′-blocked terminators — a process that entails screening large libraries of mutant DNA polymerases — has spurred the development of 3′-unblocked reversible terminators. LaserGen, Inc. was the first group to show that a small terminating group attached to the base of a 3′-unblocked nucleotide can act as an effective reversible terminator and be efficiently incorporated by wild-type DNA polymerases[31]. This led to the development of Lightning Terminators[32] (BOX 1). Helicos BioSciences has reported the development of Virtual Terminators, which are 3′-unblocked terminators with a second nucleoside analogue that acts as an inhibitor[33]. The challenge for 3′-unblocked terminators is creating the appropriate modifications to the terminating (Lightning Terminators)[32] or inhibiting (Virtual Terminators)[33] groups so that DNA synthesis is terminated after a single base addition. This is important because an unblocked 3′-OH group is the natural substrate for incorporating the next incoming nucleotide. Cleavage of only a single bond is required to remove both the terminating or inhibiting group and the fluorophore group from the nucleobase, which is a more efficient strategy than 3′-blocked terminators for restoring the nucleotide for the next CRT cycle.

Helicos BioSciences was the first group to commercialize a single-molecule sequencer, the HeliScope, which was based on the work of Quake and colleagues[34]. The HeliScope uses the single-molecule template methods shown in FIG. 1c and FIG. 1d coupled with the one-colour (Cy5 dye) CRT method shown in FIG. 2c. Incorporation of a nucleotide results in a fluorescent signal. The HeliScope also uses TIRF to image the Cy5 dye[34], the imaging output of which is shown in FIG. 2d. Harris and colleagues[14] used Cy5-12ss-dNTPs, which are earlier versions of their Virtual Terminators that lack the inhibiting group, and reported that deletion errors in homopolymeric repeat regions were the most common error type (~5% frequency) when using the primer-immobilized strategy shown in FIG. 1c. This is likely to be related to the incorporation of two or more Cy5-12ss-dNTPs in a given cycle. These errors can be greatly reduced with two-pass sequencing, which provides ~25-base consensus reads using the template-immobilized strategy shown in FIG. 1d. At the 2009 Advances in Genome Biology and Technology (AGBT) meeting, the Helicos group reported their recent progress in sequencing the

Dephasing
This occurs with step-wise addition methods when growing primers move out of synchronicity for any given cycle. Lagging strands (for example, $n - 1$ from the expected cycle) result from incomplete extension, and leading strands (for example, $n + 1$) result from the addition of multiple nucleotides or probes in a population of identical templates.

Dark nucleotides or probes
A nucleotide or probe that does not contain a fluorescent label. It can be generated from its cleavage and carry-over from the previous cycle or be hydrolysed *in situ* from its dye-labelled counterpart in the current cycle.

Total internal reflection fluorescence
A total internal reflection fluorescence imaging device produces an evanescent wave — that is, a near-field stationary excitation wave with an intensity that decreases exponentially away from the surface. This wave propagates across a boundary surface, such as a glass slide, resulting in the excitation of fluorescent molecules near (< 200 nm) or at the surface and the subsequent collection of their emission signals by a detector.

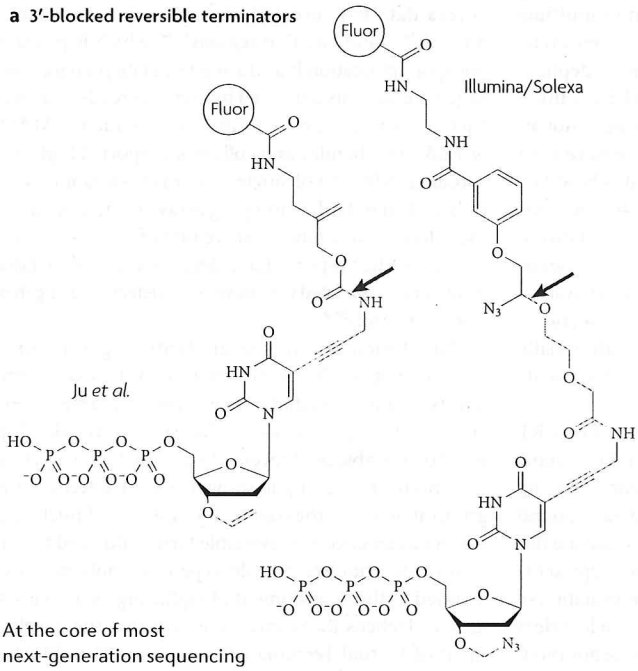
Libraries of mutant DNA polymerases
Large numbers of genetically engineered DNA polymerases can be created by either site-directed or random mutagenesis, which leads to one or more amino acid substitutions, insertions and/or deletions in the polymerase. The goal of this approach is to incorporate modified nucleotides more efficiently during the sequencing reaction.

Consensus reads
These are only useful for single-molecule techniques and are produced by sequencing the same template molecule more than once. The data are then aligned to produce a 'consensus read', reducing stochastic errors that may occur in a given sequence read.

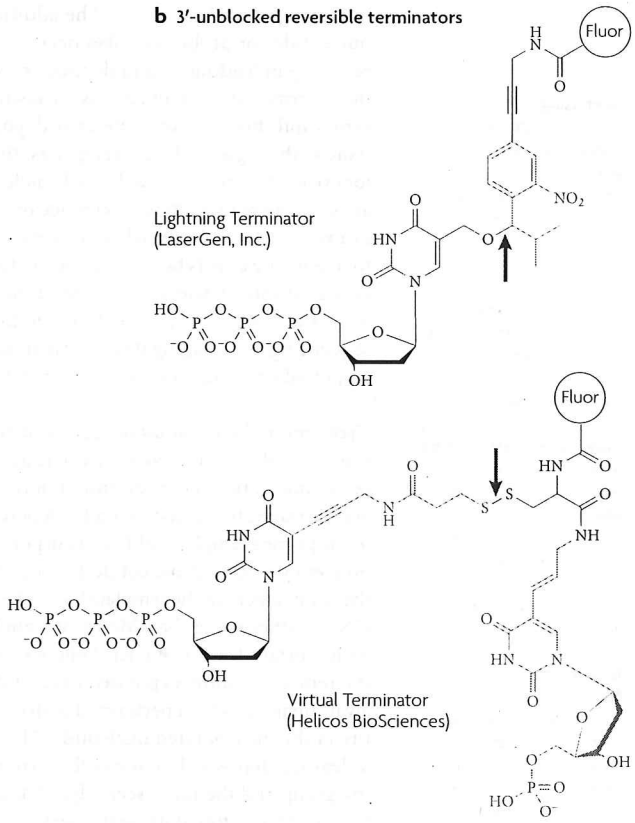
©2010 Macmillan Publishers Limited. All rights reserved

Box 1 | **Modified nucleotides used in next-generation sequencing methods**

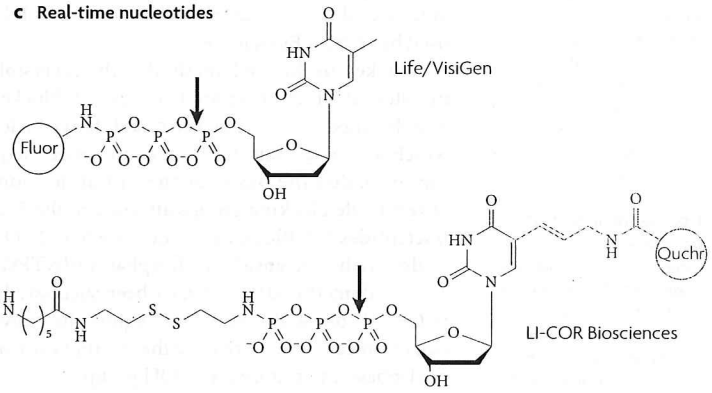
a 3′-blocked reversible terminators

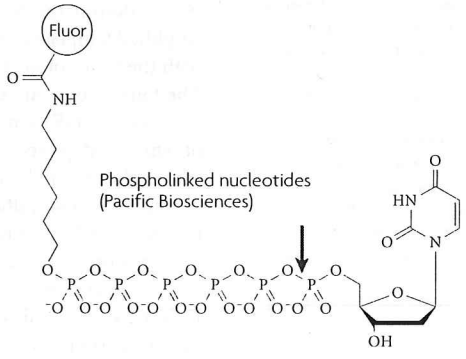


b 3′-unblocked reversible terminators



c Real-time nucleotides





At the core of most next-generation sequencing (NGS) methods is the use of dye-labelled modified nucleotides. Ideally, these nucleotides are incorporated specifically, cleaved efficiently during or following fluorescence imaging, and extended as modified or natural bases in ensuing cycles. In the figure, red chemical structures denote terminating functional groups, except in the Helicos BioSciences structure, which is characterized by an inhibitory function[33]. Arrows indicate the site of cleavage separating the fluorophore from the nucleotide, and the blue chemical structures denote residual linker structures or molecular scars that are attached to the base and accumulate with subsequent cycles. DNA synthesis is terminated by reversible terminators following the incorporation of one modified nucleotide by DNA polymerase. Two types of reversible terminators have been described: 3′-blocked terminators, which contain a cleavable group attached to the 3′-oxygen of the 2′-deoxyribose sugar, and 3′-unblocked terminators.

Several blocking groups have been described (see the figure, part **a**), including 3′-*O*-allyl[19,21,101] (Ju & colleagues, who exclusively licensed their technology to Intelligent Bio-Systems) and 3′-*O*-azidomethyl[22,23,101] (Illumina/Solexa). The blocking group attached to the 3′ end causes a bias against incorporation with DNA polymerase. Mutagenesis of DNA polymerase is required to facilitate the incorporation of 3′-blocked terminators.

3′-unblocked reversible terminators (part **b**) show more favourable enzymatic incorporation and, in some cases, can be incorporated as well as a natural nucleotide using wild-type DNA polymerases[31]. Other groups, including Church and colleagues[102] and Turcatti and colleagues[103], have described 3′-unblocked terminators that rely on steric hindrance of the bulky dye group to inhibit incorporation after the addition of the first nucleotide.

With real-time nucleotides (part **c**), the fluorophore is attached to the terminal phosphate group (Life/VisiGen[16] and Pacific Biosciences[15]) rather than the nucleobase, which also reduces bias against incorporation with DNA polymerase. In addition to labelling the terminal phosphate group, LI-COR Biosciences' nucleotides attach a quencher molecule to the base[17]. Gamma-labelled 2′-deoxyribonucleoside triphosphates (dNTPs) were first described in 1979 by Yarbrough *et al.*[104], and more recently, Kumar *et al.* described their terminally labelled polyphosphate nucleotides[105]. With the exception of LI-COR Biosciences' nucleotides, which leave the quencher group attached, natural bases are incorporated into the growing primer strand.

©2010 Macmillan Publishers Limited. All rights reserved

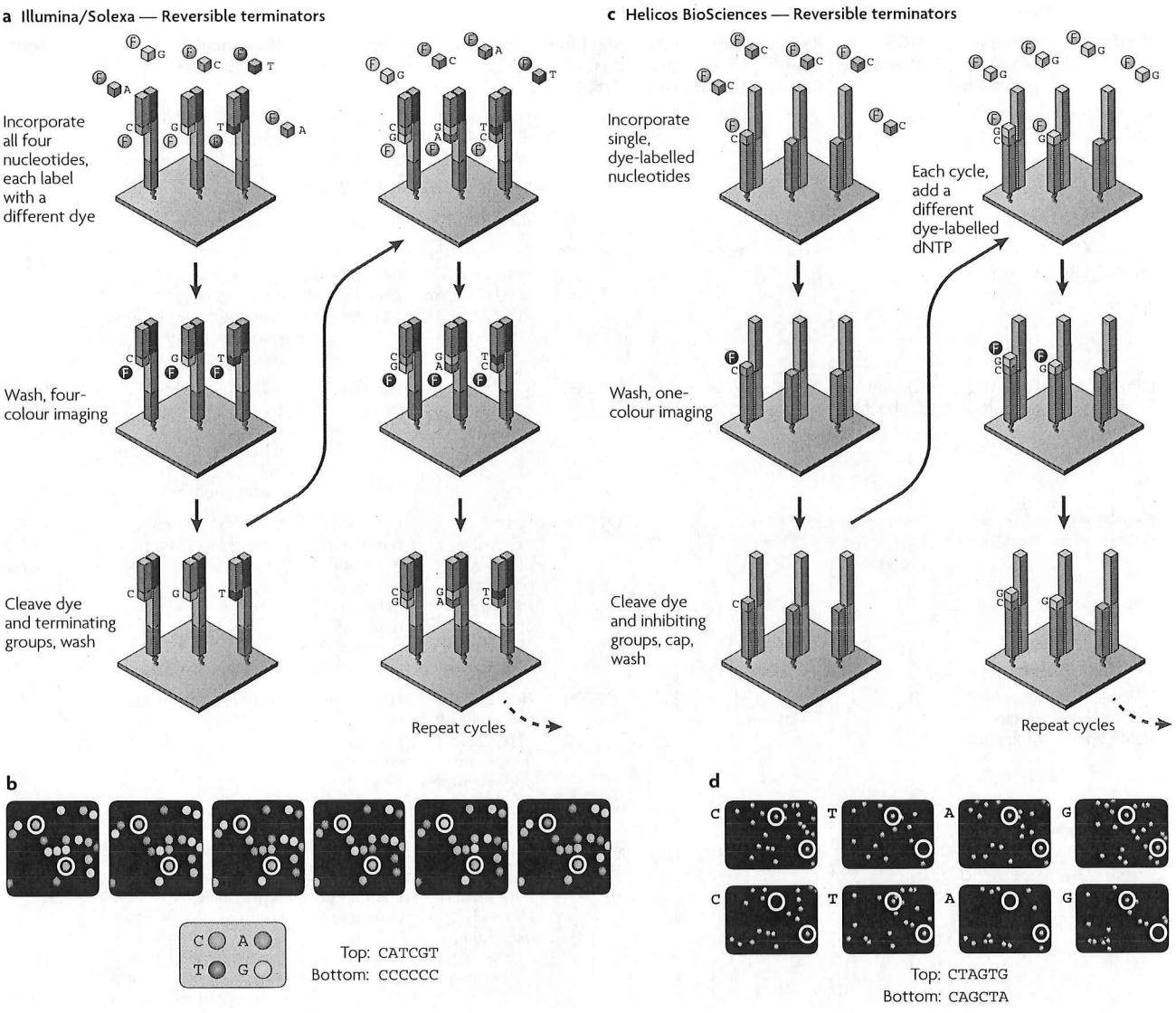


Figure 2 | **Four-colour and one-colour cyclic reversible termination methods. a** | The four-colour cyclic reversible termination (CRT) method uses Illumina/Solexa's 3′-*O*-azidomethyl reversible terminator chemistry[23,101] (BOX 1) using solid-phase-amplified template clusters (FIG. 1b, shown as single templates for illustrative purposes). Following imaging, a cleavage step removes the fluorescent dyes and regenerates the 3′-OH group using the reducing agent tris(2-carboxyethyl)phosphine (TCEP)[23]. **b** | The four-colour images highlight the sequencing data from two clonally amplified templates. **c** | Unlike Illumina/Solexa's terminators, the Helicos Virtual Terminators[33] are labelled with the same dye and dispensed individually in a predetermined order, analogous to a single-nucleotide addition method. Following total internal reflection fluorescence imaging, a cleavage step removes the fluorescent dye and inhibitory groups using TCEP to permit the addition of the next Cy5-2′-deoxyribonucleoside triphosphate (dNTP) analogue. The free sulphhydryl groups are then capped with iodoacetamide before the next nucleotide addition[33] (step not shown). **d** | The one-colour images highlight the sequencing data from two single-molecule templates.

One-base-encoded probe
An oligonucleotide sequence in which one interrogation base is associated with a particular dye (for example, A in the first position corresponds to a green dye). An example of a one-base degenerate probe set is '1-probes', which indicates that the first nucleotide is the interrogation base. The remaining bases consist of either degenerate (four possible bases) or universal bases.

Caenorhabditis elegans genome. From a single HeliScope run using only 7 of the instrument's 50 channels, approximately 2.8 Gb of high-quality data were generated in 8 days from >25-base consensus reads with 0, 1 or 2 errors. Greater than 99% coverage of the genome was reported, and for regions that showed >5-fold coverage, the consensus accuracy was 99.999% (J. W. Efcavitch, personal communication).

Sequencing by ligation. SBL is another cyclic method that differs from CRT in its use of DNA ligase[35] and either one-base-encoded probes or two-base-encoded probes. In its simplest form, a fluorescently labelled probe hybridizes to its complementary sequence adjacent to the primed template. DNA ligase is then added to join the dye-labelled probe to the primer. Non-ligated probes are washed away, followed by fluorescence

©2010 Macmillan Publishers Limited. All rights reserved

Table 1 | **Comparison of next-generation sequencing platforms**

Platform	Library/ template preparation	NGS chemistry	Read length (bases)	Run time (days)	Gb per run	Machine cost (US$)	Pros	Cons	Biological applications	Refs
Roche/454's GS FLX Titanium	Frag, MP/ emPCR	PS	330*	0.35	0.45	500,000	Longer reads improve mapping in repetitive regions; fast run times	High reagent cost; high error rates in homo-polymer repeats	Bacterial and insect genome *de novo* assemblies; medium scale (<3 Mb) exome capture; 16S in metagenomics	D. Muzny, pers. comm.
Illumina/ Solexa's GA_{II}	Frag, MP/ solid-phase	RTs	75 or 100	4‡, 9§	18‡, 35§	540,000	Currently the most widely used platform in the field	Low multiplexing capability of samples	Variant discovery by whole-genome resequencing or whole-exome capture; gene discovery in metagenomics	D. Muzny, pers. comm.
Life/APG's SOLiD 3	Frag, MP/ emPCR	Cleavable probe SBL	50	7‡, 14§	30‡, 50§	595,000	Two-base encoding provides inherent error correction	Long run times	Variant discovery by whole-genome resequencing or whole-exome capture; gene discovery in metagenomics	D. Muzny, pers. comm.
Polonator G.007	MP only/ emPCR	Non-cleavable probe SBL	26	5§	12§	170,000	Least expensive platform; open source to adapt alternative NGS chemistries	Users are required to maintain and quality control reagents; shortest NGS read lengths	Bacterial genome resequencing for variant discovery	J. Edwards, pers. comm.
Helicos BioSciences HeliScope	Frag, MP/ single molecule	RTs	32*	8‡	37‡	999,000	Non-bias representation of templates for genome and seq-based applications	High error rates compared with other reversible terminator chemistries	Seq-based methods	91
Pacific Biosciences (target release: 2010)	Frag only/ single molecule	Real-time	964*	N/A	N/A	N/A	Has the greatest potential for reads exceeding 1 kb	Highest error rates compared with other NGS chemistries	Full-length transcriptome sequencing; complements other resequencing efforts in discovering large structural variants and haplotype blocks	S. Turner, pers. comm.

*Average read-lengths. ‡Fragment run. §Mate-pair run. Frag, fragment; GA, Genome Analyzer; GS, Genome Sequencer; MP, mate-pair; N/A, not available; NGS, next-generation sequencing; PS, pyrosequencing; RT, reversible terminator; SBL, sequencing by ligation; SOLiD, support oligonucleotide ligation detection.

imaging to determine the identity of the ligated probe[36]. The cycle can be repeated either by using cleavable probes to remove the fluorescent dye and regenerate a 5′-PO_4 group for subsequent ligation cycles (FIG. 3a) or by removing and hybridizing a new primer to the template (not shown in the figure).

Shendure and colleagues[11] used the SBL method to sequence the *Escherichia coli* MG1655 genome. Mate-pair templates were prepared with four priming sites (named A1 to A4) and were amplified by emPCR (FIG. 1a). A number of one-base-encoded probes, 1-probes to 7-probes, were used. In the first SBL cycle, the A1 primer was annealed to the template, followed by the hybridization and ligation of 1-probes, four-colour imaging and removal of the entire primer–probe strand from the solid-phase-bound template. The SBL cycle was then repeated with the A1 primer but using 2-probes, and then with the A1 primer but using 3-probes, and so on. The other three primers, A2, A3 and A4, were then cycled in an analogous manner to yield six (A2 and A4) and seven (A1 and A3) base reads (albeit discontiguous) from each genomic end, making a total of 26 base reads per mate-pair template. From two instrument runs, the authors reported the production of approximately 48 million high-quality bases, which mapped to about 70% of the *E. coli* genome[11]. This SBL method is being used on the Polonator instrument.

Life/APG has commercialized their SBL platform called support oligonucleotide ligation detection (SOLiD)[37]. The method uses two-base-encoded probes, which has the primary advantage of improved accuracy in colour calling and SNV calling, the latter of which requires an adjacent valid colour change. Colour space is a unique feature of the SOLiD system. A primer is

Two-base-encoded probe
An oligonucleotide sequence in which two interrogation bases are associated with a particular dye (for example, AA, CC, GG and TT are coded with a blue dye). '1,2-probes' indicates that the first and second nucleotides are the interrogation bases. The remaining bases consist of either degenerate or universal bases.

©2010 Macmillan Publishers Limited. All rights reserved

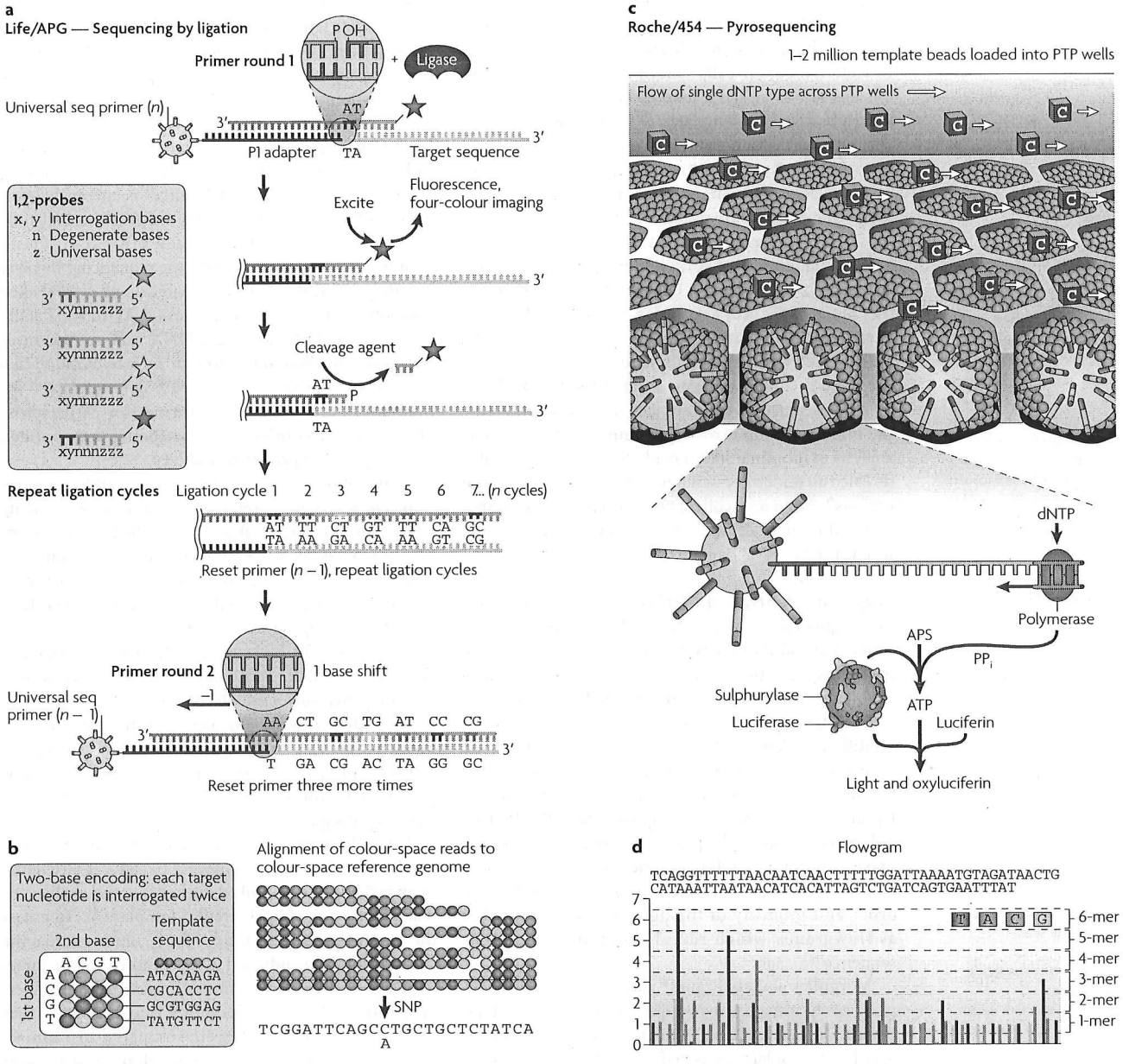


Figure 3 | **Next-generation sequencing technologies that use emulsion PCR. a** | A four-colour sequencing by ligation method using Life/APG's support oligonucleotide ligation detection (SOLiD) platform is shown. Upon the annealing of a universal primer, a library of 1,2-probes is added. Unlike polymerization, the ligation of a probe to the primer can be performed bi-directionally from either its 5′-PO_4 or 3′-OH end. Appropriate conditions enable the selective hybridization and ligation of probes to complementary positions. Following four-colour imaging, the ligated 1,2-probes are chemically cleaved with silver ions to generate a 5′-PO_4 group. The SOLiD cycle is repeated nine more times. The extended primer is then stripped and four more ligation rounds are performed, each with ten ligation cycles. The 1,2-probes are designed to interrogate the first (x) and second (y) positions adjacent to the hybridized primer, such that the 16 dinucleotides are encoded by four dyes (coloured stars). The probes also contain inosine bases (z) to reduce the complexity of the 1,2-probe library and a phosphorothiolate linkage between the fifth and six nucleotides of the probe sequence, which is cleaved with silver ions[106]. Other cleavable probe designs include RNA nucleotides[107,108] and internucleosidic phosphoramidates[107], which are cleaved by ribonucleases and acid, respectively. **b** | A two-base encoding scheme in which four dinucleotide sequences are associated with one colour (for example, AA, CC, GG and TT are coded with a blue dye). Each template base is interrogated twice and compiled into a string of colour-space data bits. The colour-space reads are aligned to a colour-space reference sequence to decode the DNA sequence. **c** | Pyrosequencing using Roche/454's Titanium platform. Following loading of the DNA-amplified beads into individual PicoTiterPlate (PTP) wells, additional beads, coupled with sulphurylase and luciferase, are added. In this example, a single type of 2′-deoxyribonucleoside triphosphate (dNTP) — cytosine — is shown flowing across the PTP wells. The fibre-optic slide is mounted in a flow chamber, enabling the delivery of sequencing reagents to the bead-packed wells. The underneath of the fibre-optic slide is directly attached to a high-resolution charge-coupled device (CCD) camera, which allows detection of the light generated from each PTP well undergoing the pyrosequencing reaction. **d** | The light generated by the enzymatic cascade is recorded as a series of peaks called a flowgram. PP_i, inorganic pyrophosphate.

©2010 Macmillan Publishers Limited. All rights reserved

hybridized to templates amplified by emPCR (FIG. 1a). The SOLiD cycle of 1,2-probe hybridization and ligation, imaging, and probe cleavage is repeated ten times to yield ten colour calls spaced in five-base intervals (FIG. 3a). The extended primer is then stripped from the solid-phase-bound templates. A second ligation round is performed with an '$n - 1$' primer, which resets the interrogation bases and the corresponding ten colour calls one position to the left. Ten ligation cycles ensue, followed by three more rounds of ligation cycles. Colour calls from the five ligation rounds are then ordered into a linear sequence (that is, the colour space) and aligned to a reference genome to decode the DNA sequence (FIG. 3b). SOLiD uses two slides per run; each can be partitioned into four or eight regions called spots. TABLE 1 depicts the current sequencing statistics of the Life/APG platform operating at the BCM-HGSC (D. Muzny, personal communication). Substitutions are the most common error type. Similar to the genome analysis of Illumina/Solexa reads, SOLiD data have also revealed an underrepresentation of AT-rich and GC-rich regions[26]. Shen and colleagues[38] recently showed that MAQ alignment of SOLiD data may be undercalling true variants.

Single-nucleotide addition: pyrosequencing. Pyrosequencing is a non-electrophoretic, bioluminescence method that measures the release of inorganic pyrophosphate by proportionally converting it into visible light using a series of enzymatic reactions[39,40] (FIG. 3c). Unlike other sequencing approaches that use modified nucleotides to terminate DNA synthesis, the pyrosequencing method manipulates DNA polymerase by the single addition of a dNTP in limiting amounts. Upon incorporation of the complementary dNTP, DNA polymerase extends the primer and pauses. DNA synthesis is reinitiated following the addition of the next complementary dNTP in the dispensing cycle. The order and intensity of the light peaks are recorded as flowgrams, which reveal the underlying DNA sequence (FIG. 3d).

Margulies and colleagues[41] described the first NGS platform to integrate pyrosequencing using their PTP device. Commercialized by Roche/454, the instrument uses DNA templates prepared by emPCR (FIG. 1a), with 1–2 million beads deposited into PTP wells. Roche/454 recently released a titanium-coated PTP design, which substantially increases read length and improves data quality by reducing crosstalk between adjacent wells containing single clonally amplified beads (T. Harkins, personal communication). Smaller beads, which have sulphurylase and luciferase attached to them to facilitate light production, are loaded into wells surrounding the template beads. Individual dNTPs are then streamed across the wells and dispensed in a predetermined sequential order. The bioluminescence is imaged with a charge-coupled device (CCD) camera (FIG. 3c). TABLE 1 shows the current sequencing statistics of the Roche/454 platform operating at the BCM-HGSC (D. Muzny, personal communication). Unlike platforms that produce shorter read lengths, the Roche/454 platform does not require the run time to be doubled for the sequencing of mate-pair templates. For homopolymeric repeats of up to six nucleotides, the number of dNTPs added is directly proportional to the light signal. Insertions are the most common error type, followed by deletions.

Real-time sequencing. The next technology method to hit the commercial sector is likely to be real-time sequencing, and Pacific Biosciences is currently leading this effort[15]. Unlike reversible terminators, real-time nucleotides do not halt the process of DNA synthesis. Simply put, the method of real-time sequencing involves imaging the continuous incorporation of dye-labelled nucleotides during DNA synthesis[42]. With the Pacific Biosciences platform, single DNA polymerase molecules are attached to the bottom surface of individual zero-mode waveguide detectors (ZMW detectors)[43] (FIG. 4a) that can obtain sequence information while phospholinked nucleotides (BOX 1) are being incorporated into the growing primer strand[15] (FIG. 4b).

Other approaches have been proposed to enhance signal-to-noise measurements in real-time sequencing using more conventional detection schemes. For example, Life/VisiGen has engineered DNA polymerases with an attached fluorescent dye that, upon incorporation of their γ-labelled nucleotides, produce an enhanced signal by fluorescence resonance energy transfer[16]. LI-COR Biosciences, whose technology was acquired by Pacific Biosciences, has been developing dye-quencher nucleotides (BOX 1), which in their native state produce low signals owing to the presence of a quencher group attached to the base. The release and diffusion of the dye-labelled pyrophosphate analogue away from the immobilized DNA polymerase produces a fluorescent signal[17].

Pacific Biosciences used the highly processive, strand-displacing φ29 DNA polymerase because it efficiently incorporates phospholinked nucleotides and enables the resequencing of closed circular templates[15]. To assess the accuracy of this method, a four-colour sequencing experiment was conducted using a known 150 bp linear template. Base calls from the real-time reads were determined from their corresponding fluorescence pulses (FIG. 4b). When the reads were compared to a known sequence, 27 errors consisting of deletions, insertions and mismatches were identified, corresponding to a read accuracy of approximately 83% (131/158). Factors that led to sequencing errors included extremely short interphase intervals between two incorporation events, and the binding and release of nucleotides in the active site before incorporation into the primer strand. Given that most errors appear as stochastic events, the authors showed that repeated sequencing of the same template molecule 15 times or more could improve the consensus read accuracy to >99%[15]. At the 2009 AGBT meeting, Pacific Biosciences reported improvements to their platform; when it was used to sequence the *E. coli* genome at 38-fold base coverage, 99.3% genome coverage was obtained. The consensus accuracy reached was >99.999% for the entire genome, with read lengths averaging 964 bases (S. Turner, personal communication).

Adjacent valid colour
A nucleotide substitution will have two colour calls, one from the 5′ position and one from the 3′ position of the dinucleotide sequence. When compared with a reference genome, base substitution in the target sequence is encoded by two specific, adjacent colours. In Figure 3b, the sequence 'CCT' is encoded as blue-yellow ('CC' = blue; 'CT' = yellow), but substituting the middle 'C' for 'A' would result in two colour changes to green-red. Any other colour sequence can be discarded as an error.

Colour space
With two-base-encoded probes, the fluorescent signal or colour obtained during imaging is associated with four dinucleotide sequences having a 5′- and 3′-base. Colour space is the sequence of overlapping dinucleotides that codes four simultaneous nucleotide sequences. Alignment with a reference genome is the most accurate method for translating colour space into a single nucleotide sequence.

Zero-mode waveguide detectors
This nanostructure device is 100 nm in diameter, which is smaller than the 532 nm and 643 nm laser wavelengths used in the Pacific Biosciences platform. Light cannot propagate through these small waveguides, hence the term zero-mode. These aluminium-clad waveguides are designed to produce an evanescent wave (see the 'total internal reflection fluorescence' glossary term) that substantially reduces the observation volume at the surface of the polymerase reaction down to the zeptolitre range (10^{-21} l). This provides an advantage for the polymerization reaction, which can be performed at higher dye-labelled nucleotide concentrations.

©2010 Macmillan Publishers Limited. All rights reserved

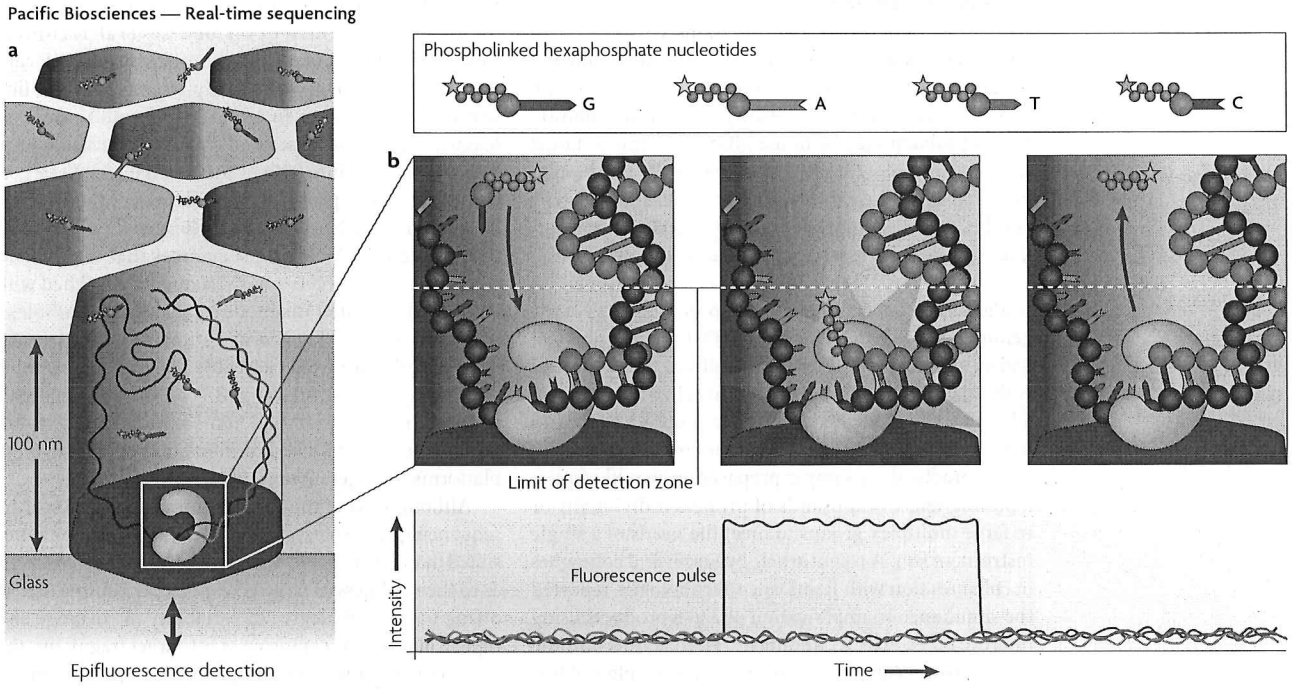


Figure 4 | **Real-time sequencing.** Pacific Biosciences' four-colour real-time sequencing method is shown. **a** | The zero-mode waveguide (ZMW) design reduces the observation volume, therefore reducing the number of stray fluorescently labelled molecules that enter the detection layer for a given period. These ZMW detectors address the dilemma that DNA polymerases perform optimally when fluorescently labelled nucleotides are present in the micromolar concentration range, whereas most single-molecule detection methods perform optimally when fluorescent species are in the pico- to nanomolar concentration range[42]. **b** | The residence time of phospholinked nucleotides in the active site is governed by the rate of catalysis and is usually on the millisecond scale. This corresponds to a recorded fluorescence pulse, because only the bound, dye-labelled nucleotide occupies the ZMW detection zone on this timescale. The released, dye-labelled pentaphosphate by-product quickly diffuses away, dropping the fluorescence signal to background levels. Translocation of the template marks the interphase period before binding and incorporation of the next incoming phospholinked nucleotide.

Genome alignment and assembly

After NGS reads have been generated, they are aligned to a known reference sequence or assembled *de novo*[8,44,45]. The decision to use either strategy is based on the intended biological application as well as cost, effort and time considerations. For example, identifying and cataloguing genetic variation in multiple strains of highly related genomes, such as those found in specific species of bacteria[46–51], *C. elegans*[25,30,38] and *Arabidopsis thaliana*[52], can be accomplished by aligning NGS reads to their reference genomes. This approach is substantially cheaper and faster than Sanger sequencing. SNVs can be readily identified, although in many cases, validation of these findings is required.

There are limitations to the alignment approach, such as placing reads within repetitive regions in the reference genome or in corresponding regions that may not exist in the reference genome[28]; the latter situation may result from gaps in the reference genome or the presence of structural variants (SVs) in the genome being analysed. Mate-pair reads can resolve the correct genome assignment for some repetitive regions as long as one read in the pair is unique to the genome. A study by Egholm, Snyder and colleagues[53] showed that Roche/454 read data (averaging 258 bases), derived from 3-kb sized mate-pair genomic libraries, could capture a larger fraction of SVs in the human genome, although this approach still identified fewer SVs than traditional fosmid-end sequencing approaches[54].

De novo assemblies have been reported for bacterial genomes and mammalian bacterial artificial chromosomes[55–59], but substantial challenges exist for their application to human genomes. A reasonable strategy for improving the quality of the alignment or assembly has been to increase the read coverage. An article by Frazer and colleagues[26] challenges this approach by reporting systematic variability in local sequence coverage that was specific to different human genomic regions for the Roche/454, Illumina/Solexa and Life/APG platforms (other commercially available platforms were not evaluated). Because each NGS platform examined produced a uniquely reproducible pattern of variable sequence coverage, the mixing of different NGS read types in the alignment or assembly may remedy this shortcoming. Recently it was reported that mixing Roche/454 and Illumina/Solexa read data resulted in improved *de novo* assemblies of microbial genomes compared with assemblies based on data from either platform alone[60,61].

Fluorescence resonance energy transfer
This is generally a system that consists of two fluorescent dyes, one being a donor dye (a bluer fluorophore) and the other an acceptor dye (a redder fluorophore). When the two dye molecules are brought into close proximity (usually ≤30 nm), the energy from the excited donor dye is transferred to the acceptor dye, increasing its emission intensity signal.

Structural variants
All sequence variants other than single-nucleotide variants, including block substitutions, insertions or deletions, inversions, segmental duplications and copy-number differences.

©2010 Macmillan Publishers Limited. All rights reserved

Genome enrichment

Despite the substantial cost reductions associated with NGS technologies in comparison with the automated Sanger method[62], whole-genome sequencing is still an expensive endeavour (see below). An interim solution to this problem may be to use NGS platforms to target specific regions of interest. This strategy can be used to examine all of the exons in the genome, specific gene families that constitute known drug targets or megabase-size regions that are implicated in disease[63] or pharmacogenetics effects[64] through genome-wide association studies. The concept of targeting specific regions of the genome is well established, with PCR being the most widely used method, albeit on a small scale. PCR coupled with Sanger sequencing is suitably matched for analysing a handful of candidate genes[65], but coupling PCR with high-throughput NGS platforms for targeting strategies is not practical, as sample preparation would require handling tens of thousands of primers individually or in large multiplex groups to meet the needs of a single instrument run. A recent article by Frazer and colleagues in collaboration with RainDance Technologies reported the simultaneous amplification of 3,976 products using microdroplet PCR technology[66]. Here, a microfluidic device creates aqueous picolitre-volume droplets of forward- and reverse-targeting primers in an oil solution. The primer droplets that are targeted to different regions of the genome merge with separate picolitre droplets that contain fragmented genomic DNA and associated PCR reagents, and these mixed droplets are thermal cycled in a single tube (FIG. 5a). The authors reported an 84% capture efficiency with 90% of the targeted bases showing uniform coverage using the microdroplet PCR method sequenced with either the Roche/454 or Illumina/Solexa platform[66].

Custom-designed oligonucleotide microarrays and solution-based hybridization strategies have also been used for targeting regions of interest. For example, Roche/NimbleGen[67] oligonucleotide microarrays have been designed to provide complementary sequences for solid-phase hybridization to enrich for exons[68,69] or contiguous stretches of genomic regions[68,70] (FIG. 5b). The BCM-HGSC[68] and Cold Spring Harbor Laboratory[69] groups reported capture efficiencies of 65–77% (Roche/454 platform) and 53% (Illumina/Solexa platform), respectively, with targeted exons being covered by at least one NGS read. Recently, the BCM-HGSC group reported >90% capture efficiency with at least 10× base coverage (Roche/454 and Life/APG platforms) by chip optimization that decreases the number of probes in higher coverage regions while increasing those in low coverage regions (R. A. Gibbs, personal communication).

Other groups have captured specific genomic regions by using solution-based hybridization methods, such as molecular inversion probes (MIPs)[71–73] and biotinylated RNA capture sequences[74] (FIG. 5c). Shendure, Church and colleagues[71] were the first group to use MIPs in exon targeting by designing specific oligonucleotide ends that flank the exons of interest. However, in duplicate experiments only 20% of the targets were captured and even fewer exon regions (11%) were found in both data sets (Illumina/Solexa platform)[71]. Mindrinos, Davis and colleagues[72] recently described several technical improvements that increased the efficiency of MIP capture to >90%, with approximately 70% of targets falling within a 10-fold range in 485 exons. Shendure and colleagues[73] have also described improvements that enabled them to capture 91% of their initial 55,000 targets (Illumina/Solexa platform). An alternative method has been reported by Nusbaum and colleagues[74], who created biotinylated RNA capture sequences that are hybridized to genomic targets and subsequently enriched with streptavidin-coated magnetic beads (Illumina/Solexa platform) (FIG. 5c). This group estimated a capture efficiency of 60% for exons and 80% for genomic regions. The 1000 Genomes Project and The Exome Project have adopted microarray-based and solution-based strategies that use Roche/454, Illumina/Solexa and Life/APG platforms to target regions of interest.

Although any of the NGS platforms can perform the sequencing phase of targeted capture, some may be better suited than others (TABLE 1). Initial reports raised concerns as to the readiness of targeted capture for routine use[75,76] owing to the high degree of variability in coverage and representation, the capturing of off-target fragments, the requirement for microgram quantities of starting material and the high cost for microarrays, but improvements are being made to address these limitations[76].

Next-generation sequencing applications

The production of large numbers of low-cost reads makes the NGS platforms described above useful for many applications. These include variant discovery by resequencing targeted regions of interest or whole genomes, *de novo* assemblies of bacterial and lower eukaryotic genomes, cataloguing the transcriptomes of cells, tissues and organisms (RNA–seq)[5], genome-wide profiling of epigenetic marks and chromatin structure using other seq-based methods (ChIP–seq, methyl–seq and DNase–seq)[4], and species classification and/or gene discovery by metagenomics studies[77]. With this many applications, which platform is best suited for a given biological experiment? For example, the Illumina/Solexa and Life/APG platforms are well suited for variant discovery by resequencing human genomes because gigantic volumes of high-quality bases are produced per run (TABLE 1). Furthermore, the Helicos BioSciences platform is well suited for applications that demand quantitative information in RNA–seq[78] or direct RNA sequencing, as it sequences RNA templates directly without the need to convert them into cDNAs[79]. TABLE 1 provides an overview of NGS technologies, instrument performance and cost, pros and cons, and recommendations for biological applications; however, the rapid pace of technological advances in the field could change this information in the near future. Readers are directed to several excellent reviews on RNA–seq[5], ChIP–seq[80] and metagenomics[77].

In the following section, I highlight recent advances in sequencing personal genomes using NGS, as demand in this area is driving rapid technological developments and competitive pricing.

1000 Genomes Project
A project aimed at discovering rare sequence variants with minor allele frequencies of 1% in normal genomes derived from HapMap samples.

The Exome Project
A project aimed at developing and validating cost-effective, high-throughput technologies for resequencing all of the protein-coding regions of the human genome.

Metagenomics
The study of communities of mixed microbial genomes that reside in animals, plants and environmental niches. Samples are collected and analysed without the need to culture isolated microbes in the laboratory. The Human Microbiome Project aims to characterize a reference set of microbial genomes from different habitats within the human body, including nasal, oral, skin, gastrointestinal and urogenital regions, and to determine how changes in the human microbiome affect health and disease.

©2010 Macmillan Publishers Limited. All rights reserved

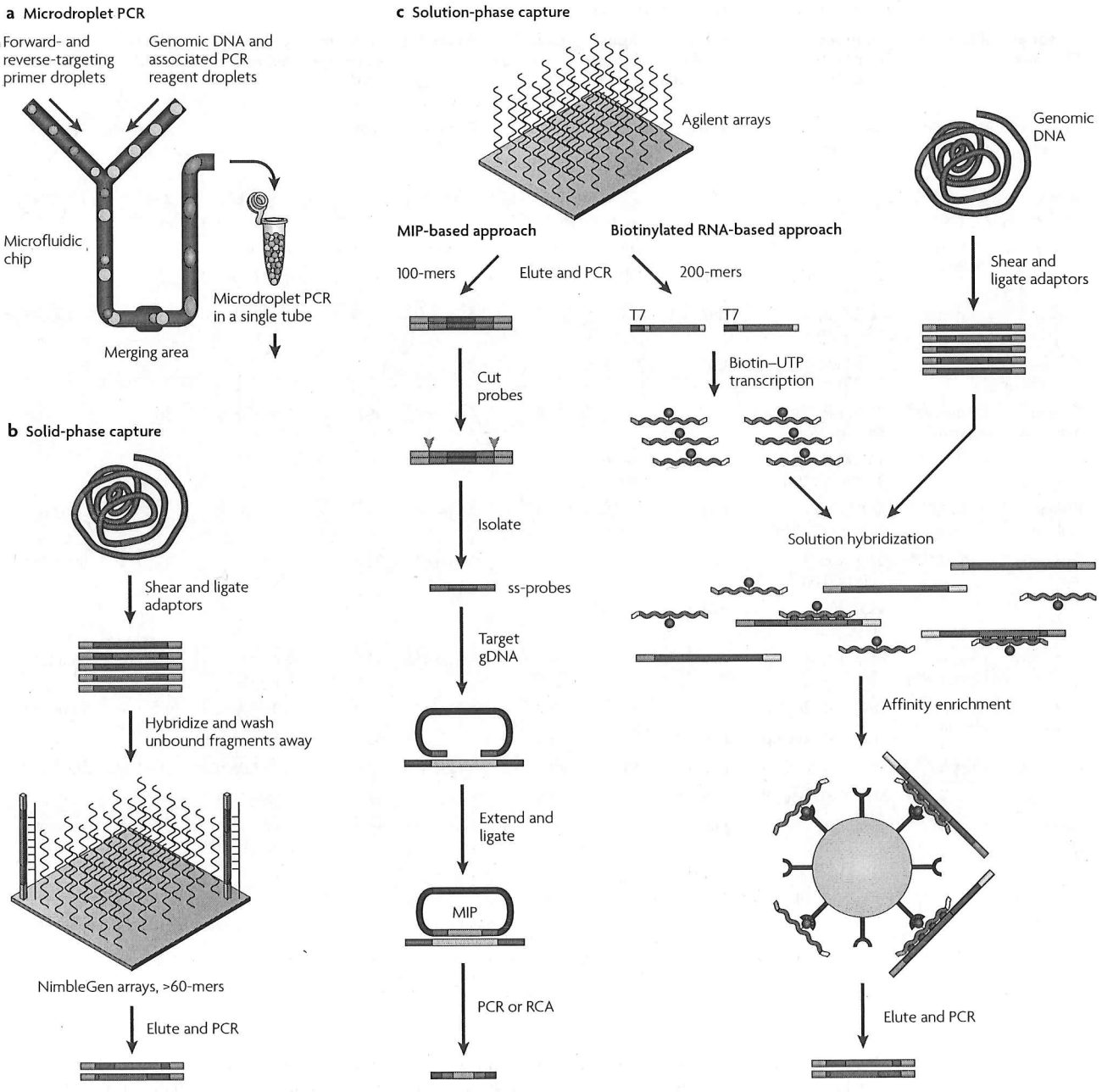


Figure 5 | **Targeted capture scheme. a** | A microfluidic nozzle creates aqueous droplets (8 pl) of forward- and reverse-targeting primers in oil (not shown), which are dispensed into a channel in a microfluidic chip. Fragmented genomic DNA and associated PCR reagents are separately dispensed as aqueous microdroplets (14 pl) and paired together with the primer droplets at a 1/1 ratio. Paired droplets pass through a merging area for coalescence into PCR-competent microdroplets, which are collected into a single tube for thermal cycling. **b** | Solid-phase capture methods ligate adaptor sequences to fragmented genomic DNA before microarray hybridization. Following stringent washing to remove non-selected genomic fragments, the enriched targets are eluted, PCR amplified and sequenced. Adaptor sequences can also be designed specifically for a given next-generation sequencing platform, which eliminates the PCR step. **c** | Solution-phase methods use custom-designed microarrays from Agilent Technologies to synthesize long, target-specific oligonucleotides, which are cleaved as a pool of probes from the support and eluted into a single tube. In the molecular inversion probe (MIP)-based approach, single-stranded probes (ss-probes) are hybridized to their corresponding templates. DNA polymerase extends the 3′-end of the probes across the exon region, and a ligation step creates closed circular DNAs. The biotinylated RNA-based approach introduces a T7 promoter sequence, which allows the incorporation of biotin-labelled uridine-5′-triphosphate (UTP) into the probe sequence by *in vitro* transcription. Following solution hybridization with genomic targets, the RNA–DNA hybrids are enriched by their binding to streptavidin-coated magnetic beads. RCA, rolling circle amplification.

©2010 Macmillan Publishers Limited. All rights reserved

Table 2 | **Sequencing statistics on personal genome projects**

Personal Genome	Platform	Genomic template libraries	No. of reads (millions)	Read length (bases)	Base coverage (fold)	Assembly	Genome coverage (%)*	SNVs in millions (alignment tool)	No. of runs	Estimated cost (US$)
J. Craig Venter	Automated Sanger	MP from BACs, fosmids & plasmids	31.9	800	7.5	*De novo*	N/A	3.21	>340,000	70,000,000
James D. Watson	Roche/454	Frag: 500 bp	93.2‡	250§	7.4	Aligned*	95‖	3.32 (BLAT)	234	1,000,000¶
Yoruban male (NA18507)	Illumina/ Solexa	93% MP: 200 bp	3,410‡	35	40.6	Aligned*	99.9	3.83 (MAQ)	40	250,000¶
		7% MP: 1.8 kb	271	35				4.14 (ELAND)		
Han Chinese male	Illumina/ Solexa	66% Frag: 150–250 bp	1,921‡	35	36	Aligned*	99.9	3.07 (SOAP)	35	500,000¶
		34% MP: 135 bp & 440 bp	1,029	35						
Korean male (AK1)	Illumina/ Solexa	21% Frag: 130 bp & 440 bp	393‡	36	27.8	Aligned*	99.8	3.45 (GSNAP)	30	200,000¶
		79% MP: 130 bp, 390 bp & 2.7 kb	1,156	36, 88, 106						
Korean male (SJK)	Illumina/ Solexa	MP: 100 bp, 200 bp & 300 bp	1,647‡	35, 74	29.0	Aligned*	99.9	3.44 (MAQ)	15	250,000¶,#
Yoruban male (NA18507)	Life/APG	9% Frag: 100–500 bp	211‡	50	17.9	Aligned*	98.6	3.87 (Corona-lite)	9.5	60,000¶,**
		91% MP: 600–3,500 bp	2,075‡	25, 50						
Stephen R. Quake	Helicos BioSciences	Frag: 100–500 bp	2,725‡	32§	28	Aligned*	90	2.81 (IndexDP)	4	48,000¶
AML female	Illumina/ Solexa	Frag: 150–200 bp‡‡	2,730‡,‡‡	32	32.7	Aligned*	91	3.81‡‡ (MAQ)	98	1,600,000¶¶
		Frag: 150–200 bp§§	1,081‡,§§	35	13.9		83	2.92§§ (MAQ)	34	
AML male	Illumina/ Solexa	MP: 200–250 bp‡‡	1,620‡,‡‡	35	23.3	Aligned*	98.5	3.46‡‡ (MAQ)	16.5	500,000¶¶
		MP: 200–250 bp§§	1,351‡,§§	50	21.3		97.4	3.45§§ (MAQ)	13.1	
James R. Lupski CMT male	Life/APG	16% Frag: 100–500 bp	238‡	35	29.6	Aligned*	99.8	3.42 (Corona-lite)	3	75,000¶,¶¶
		84% MP: 600–3,500 bp	1,211‡	25, 50						

*A minimum of one read aligning to the National Center for Biotechnology Information build 36 reference genome. ‡Mappable reads for aligned assemblies. §Average read-length. ‖D. Wheeler, personal communication. ¶Reagent cost only. #S.-M. Ahn, personal communication. **K. McKernan, personal communication. ‡‡Tumour sample. §§Normal sample. ¶¶Tumour & normal samples: reagent, instrument, labour, bioinformatics and data storage cost, E. Mardis, personal communication. ¶¶R. Gibbs, personal communication. AML, acute myeloid leukaemia; BAC, bacterial artificial chromosome; CMT, Charcot–Marie–Tooth disease; Frag, fragment; MP, mate-pair; N/A, not available; SNV, single-nucleotide variant.

Personal genomes. Human genome studies aim to catalogue SNVs and SVs and their association to phenotypic differences, with the eventual goal of personalized genomics for medical purposes. In 2004, the International Human Genome Sequencing Consortium published the first, and still only, finished-grade human reference genome (currently National Center for Biotechnology Information (NCBI) build 36)[1]. Its cost was estimated at US$300 million. In October 2007, Venter and colleagues[81] described the genome sequence of J. Craig Venter using a whole-genome shotgun approach coupled with automated Sanger sequencing (TABLE 2). When the Venter genome was compared with the reference genome, 3.2 million SNVs were identified. In addition, there were over 900,000 SVs, which altogether accounted for more variant bases than the SNVs.

The first group to apply NGS to whole human genomes was Roche/454 in collaboration with Gibbs and colleagues at the BCM-HGSC, who reported the diploid genome of James D. Watson[82] (TABLE 2). When the Watson genome was compared with the reference genome, roughly 3.3 million SNVs were identified, but far fewer SVs were found than in the study by Venter and colleagues. This is important because undiscovered SVs could account for a substantial fraction of the total number of sequence variants, many of which could be potentially causative in disease[83–86]. Within the past year, five additional human genomes have been described[23,87–91], one of which was sequenced on two different NGS platforms (TABLE 2). As with the Watson genome, far fewer SVs were reported than in the study by Venter and colleagues.

©2010 Macmillan Publishers Limited. All rights reserved

The Cancer Genome Atlas
A project aimed at discovering single-nucleotide variants and structural variants that are associated with major cancers, such as brain cancer (glioblastoma multiforme), lung cancer (squamous carcinoma) and ovarian cancer (serous cystadenocarcinoma).

Personal Genome Project
A project aimed at providing open access to human genome sequences from volunteers and to develop tools for interpreting this information and correlating it with related personal medical information.

Personal genomics is also being applied to the study of disease. For example, Mardis and colleagues[29,92] have reported the sequencing of two acute myeloid leukaemia cancer genomes using the Illumina/Solexa platform, and both studies identified somatic mutations that may be associated with the disease. Gibbs and colleagues have recently described the elucidation of both allelic variants in a family with a recessive form of Charcot–Marie–Tooth disease using the Life/APG platform[93] (TABLE 2).

Several projects aimed at sequencing more individuals, including The Cancer Genome Atlas[94] and the 1000 Genomes Project, are also using the Illumina/Solexa and Life/454 platforms to sequence whole genomes. Complete Genomics (see below) recently described the first genome sequence from a Caucasian male (PGP1) enrolled in the Personal Genome Project. These projects should lead to a substantial increase in the number of personal genomes sequenced in the near future.

In comparison with automated Sanger sequencing, NGS platforms have dramatically increased throughput and substantially lowered expenditure, with several groups reporting reagent costs of below US$100,000 (TABLE 2). However, there is variability among and within NGS platforms in terms of template size and construct, read-length, throughput, and base and genome coverage (TABLE 2), and such variability makes it difficult to assess the quality (that is, the base accuracy, genome coverage and genome continuity) of genomes based on cost considerations. In June 2009, Illumina announced a personal genome sequencing service that provides 30-fold base coverage for the price of US$48,000. Complete Genomics offers a similar service with 40-fold coverage priced at US$5,000, based on a business model that is reliant on huge customer volume. Recently, Drmanac, Ballinger and colleagues sequenced the whole genomes of three individuals, including PGP1, using a modified non-cleavable probe SBL method known as combinatorial probe anchor ligation (cPAL). The number of SNVs identified was consistent with those reported in TABLE 1 for other genomes, and the group reported an amortized cost of US$4,400 in reagents[95]. Although impressive, it is unclear whether this price will be sustainable as a retail service[96] or whether the Illumina or Complete Genomics business models will see long-term profits. Achieving these cost advantages may also come with a substantial trade-off, namely the production of lower quality genomes that inadequately capture the extent of SVs[83–86]. Closing the gap between $10,000 and $1,000 will be the greatest challenge for current technology developers, and the $1,000 genome might result from as-yet-undeveloped innovations. A timetable for the $1,000 draft genome is difficult to predict, and even more uncertain is the delivery of a high-quality, finished-grade personal genome.

Closing remarks

Since 2004, the National Human Genome Research Institute has awarded more than $100 million for the development of NGS technologies[62], and these awards have facilitated much of the progress in NGS to date, as well as several commercial developments. Many companies, including IBM, Oxford Nanopore Technologies, Intelligent Bio-Systems, LaserGen, Inc. and NABsys, have NGS technologies at various stages of development and commercialization.

The production of billions of NGS reads has also challenged the infrastructure of existing information technology systems in terms of data transfer, storage and quality control, computational analysis to align or assembly read data, and laboratory information management systems for sample tracking and process management. Advances in bioinformatics are ongoing, and improvements are needed if these systems are to keep pace with the continuing developments in NGS technologies. It is possible that the costs associated with downstream data handling and analysis could match or surpass the data-production costs described in this Review.

NGS technologies have an impressive range of applications, and more are being developed. In addition to the applications described above, NGS technologies are being used to characterize the evolutionary relationships of ancient genomes[97,98] and to elucidate the role of non-coding RNAs in health and disease[99,100]. In the not too distant future, it is foreseeable that NGS technologies could be used to obtain high-quality sequence data from a genome isolated from a single cell, which would be a substantial breakthrough, particularly for cancer genomics. For this to occur, advances will be needed in techniques for efficiently isolating intact long DNA molecules and in NGS methods for accurately reading the sequence content of these molecules. The field of NGS development and applications is a fast-moving area of research, which makes this an exciting time for genomic studies.

1. International Human Genome Consortium. Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931–945 (2004).
2. Metzker, M. L. Emerging technologies in DNA sequencing. *Genome Res.* **15**, 1767–1776 (2005).
3. Hutchison, C. A. III. DNA sequencing: bench to bedside and beyond. *Nucleic Acids Res.* **35**, 6227–6237 (2007).
4. Wold, B. & Myers, R. M. Sequence census methods for functional genomics. *Nature Methods* **5**, 19–21 (2008).
5. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Rev. Genet.* **10**, 57–63 (2009).
The Review provides a comprehensive overview of recent advances and challenges in techniques that are used in transcriptome profiling methods that use NGS technologies (RNA–seq).
6. Branton, D. *et al.* The potential and challenges of nanopore sequencing. *Nature Biotech.* **26**, 1146–1153 (2008).
An excellent review of the current state of nanopore sequencing that highlights recent accomplishments and remaining challenges in the field.
7. Fan, J.-B., Chee, M. S. & Gunderson, K. L. Highly parallel genomic assays. *Nature Rev. Genet.* **7**, 632–644 (2006).
8. Pop, M. & Salzberg, S. L. Bioinformatics challenges of new sequencing technology. *Trends Genet.* **24**, 142–149 (2008).
9. Dressman, D., Yan, H., Traverso, G., Kinzler, K. W. & Vogelstein, B. Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations. *Proc. Natl. Acad. Sci. USA* **100**, 8817–8822 (2003).
10. Fedurco, M., Romieu, A., Williams, S., Lawrence, I. & Turcatti, G. BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies. *Nucleic Acids Res.* **34**, e22 (2006).
11. Shendure, J. *et al.* Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* **309**, 1728–1732 (2005).
This paper describes the development of the non-cleavable SBL method and shows its feasibility by sequencing the *E. coli* genome. The prototype described led to the development of the Polonator instrument.
12. Kim, J. B. *et al.* Polony multiplex analysis of gene expression (PMAGE) in mouse hypertrophic cardiomyopathy. *Science* **316**, 1481–1484 (2007).
13. Leamon, J. H. A massively parallel PicoTiterPlate based platform for discrete picoliter-scale polymerase chain reactions. *Electrophoresis* **24**, 3769–3777 (2003).

©2010 Macmillan Publishers Limited. All rights reserved

14. Harris, T. D. *et al.* Single-molecule DNA sequencing of a viral genome. *Science* **320**, 106–109 (2008).
Developers from Helicos BioSciences and colleagues describe the development of the first single-molecule sequencing method using reversible terminators and demonstrate the technology by sequencing the M13 genome.
15. Eid, J. *et al.* Real-time DNA sequencing from single polymerase molecules. *Science* **323**, 133–138 (2009).
The authors describe the development of a real-time sequencing method using their ZMW detection system and demonstrate its feasibility by sequencing synthetic templates.
16. Hardin, S., Gao, X., Briggs, J., Willson, R. & Tu, S.-C. Methods for real-time single molecule sequence determination. US Patent 7,329,492 (2000).
17. Williams, J. G. K. System and methods for nucleic acid sequencing of single molecules by polymerase synthesis. US Patent 6,255,083 (1998).
18. Erlich, Y., Mitra, P. P., delaBastide, M., McCombie, W. R. & Hannon, G. J. Alta-Cyclic: a self-optimizing base caller for next-generation sequencing. *Nature Methods* **5**, 679–682 (2008).
19. Metzker, M. L. *et al.* Termination of DNA synthesis by novel 3′-modified deoxyribonucleoside triphosphates. *Nucleic Acids Res.* **22**, 4259–4267 (1994).
20. Canard, B. & Sarfati, R. DNA polymerase fluorescent substrates with reversible 3′-tags. *Gene* **148**, 1–6 (1994).
21. Ju, J. *et al.* Four-color DNA sequencing by synthesis using cleavable fluorescent nucleotide reversible terminators. *Proc. Natl. Acad. Sci. USA* **103**, 19635–19640 (2006).
22. Guo, J. *et al.* Four-color DNA sequencing with 3′-*O*-modified nucleotide reversible terminators and chemically cleavable fluorescent dideoxynucleotides. *Proc. Natl Acad. Sci. USA* **105**, 9145–9150 (2008).
23. Bentley, D. R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53–59 (2008).
Developers from Illumina/Solexa and colleagues report details on their reversible terminator platform and demonstrate the technology by sequencing a flow-sorted X chromosome and the genome from a Yoruban male.
24. Dohm, J. C., Lottaz, C., Borodina, T. & Himmelbauer, H. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res.* **36**, e105 (2008).
25. Hillier, L. W. *et al.* Whole-genome sequencing and variant discovery in *C. elegans*. *Nature Methods* **5**, 183–188 (2008).
26. Harismendy, O. *et al.* Evaluation of next generation sequencing platforms for population targeted sequencing studies. *Genome Biol.* **10**, R32 (2009).
27. Li, H., Ruan, J. & Durbin, R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.* **18**, 1851–1858 (2008).
28. Frazer, K. A., Murray, S. S., Schork, N. J. & Topol, E. J. Human genetic variation and its contribution to complex traits. *Nature Rev. Genet.* **10**, 241–251 (2009).
29. Ley, T. J. *et al.* DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature* **456**, 66–72 (2008).
30. Sarin, S., Prabhu, S., O'Meara, M. M., Pe'er, I. & Hobert, O. *Caenorhabditis elegans* mutant allele identification by whole-genome sequencing. *Nature Methods* **5**, 865–867 (2008).
31. Wu, W. *et al.* Termination of DNA synthesis by N^6-alkylated, not 3′-*O*-alkylated, photocleavable 2′-deoxyadenosine triphosphates. *Nucleic Acid Res.* **35**, 6339–6349 (2007).
32. Wu, W., Litosh, V. A., Stupi, B. P. & Metzker, M. L. Photocleavable labeled nucleotides and nucleosides and methods for their use in DNA sequencing. US Patent Application 11/567,189 (2009).
33. Bowers, J. *et al.* Virtual terminator nucleotides for next-generation DNA sequencing. *Nature Methods* **6**, 593–595 (2009).
34. Braslavsky, I., Hebert, B., Kartalov, E. & Quake, S. R. Sequence information can be obtained from single DNA molecules. *Proc. Natl. Acad. Sci. USA* **100**, 3960–3964 (2003).
35. Tomkinson, A. E., Vijayakumar, S., Pascal, J. M. & Ellenberger, T. DNA ligases: structure, reaction mechanism, and function. *Chem. Rev.* **106**, 687–699 (2006).
36. Landegren, U., Kaiser, R., Sanders, J. & Hood, L. A ligase-mediated gene detection technique. *Science* **241**, 1077–1080 (1988).
37. Valouev, A. *et al.* A high-resolution, nucleosome position map of *C. elegans* reveals a lack of universal sequence-dictated positioning. *Genome Res.* **18**, 1051–1063 (2008).
This paper describes Life/APG's SBL method, which uses cleavable two-base-encoded probes on the SOLiD platform. The authors demonstrate the technology through the application of genome-wide nucleosome mapping in *C. elegans*.
38. Shen, Y., Sarin, S., Liu, Y., Hobert, O. & Pe'er, I. Comparing platforms for *C. elegans* mutant identification using high-throughput whole-genome sequencing. *PLoS ONE* **3**, e4012 (2008).
39. Ronaghi, M., Uhlén, M. & Nyrén, P. A sequencing method based on real-time pyrophosphate. *Science* **281**, 363–365 (1998).
40. Ronaghi, M., Karamohamed, S., Pettersson, B., Uhlén, M. & Nyrén, P. Real-time DNA sequencing using detection of pyrophosphate release. *Anal. Biochem.* **242**, 84–89 (1996).
41. Margulies, M. *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, 376–380 (2005).
The authors describe the development of the first NGS technology using the pyrosequencing method and demonstrate its feasibility through the sequencing and *de novo* assembly of the *Mycoplasma genitalium* genome.
42. Metzker, M. L. Sequencing in real time. *Nature Biotech.* **27**, 150–151 (2009).
43. Levene, M. J. *et al.* Zero-mode waveguides for single-molecule analysis at high concentrations. *Science* **299**, 682–686 (2003).
44. Trapnell, C. & Salzberg, S. L. How to map billions of short reads onto genomes. *Nature Biotech.* **27**, 455–457 (2009).
45. Chaisson, M. J., Brinza, D. & Pevzner, P. A. *De novo* fragment assembly with short mate-paired reads: does the read length matter? *Genome Res.* **19**, 336–346 (2009).
46. Hofreuter, D. *et al.* Unique features of a highly pathogenic *Campylobacter jejuni* strain. *Infect. Immun.* **74**, 4694–4707 (2006).
47. Holt, K. E. *et al.* High-throughput sequencing provides insights into genome variation and evolution in *Salmonella* Typhi. *Nature Genet.* **40**, 987–993 (2008).
48. Srivatsan, A. *et al.* High-precision, whole-genome sequencing of laboratory strains facilitates genetic studies. *PLoS Genet.* **4**, e1000139 (2008).
49. Suchland, R. J. *et al.* Identification of concomitant infection with *Chlamydia trachomatis* IncA-negative mutant and wild-type strains by genomic, transcriptional, and biological characterizations. *Infect. Immun.* **76**, 5438–5446 (2008).
50. Nusbaum, C. *et al.* Sensitive, specific polymorphism discovery in bacteria using massively parallel sequencing. *Nature Methods* **6**, 67–69 (2009).
51. Moran, N. A., McLaughlin, H. J. & Sorek, R. The dynamics and time scale of ongoing genomic erosion in symbiotic bacteria. *Science* **323**, 379–382 (2009).
52. Ossowski, S. *et al.* Sequencing of natural strains of *Arabidopsis thaliana* with short reads. *Genome Res.* **18**, 2024–2033 (2008).
53. Korbel, J. O. *et al.* Paired-end mapping reveals extensive structural variation in the human genome. *Science* **318**, 420–426 (2007).
54. Kidd, J. M. *et al.* Mapping and sequencing of structural variation from eight human genomes. *Nature* **453**, 56–64 (2008).
55. Warren, R. L., Sutton, G. G., Jones, S. J. M. & Holt, R. A. Assembling millions of short DNA sequences using SSAKE. *Bioinformatics* **23**, 500–501 (2007).
56. Chaisson, M. J. & Pevzner, P. A. Short read fragment assembly of bacterial genomes. *Genome Res.* **18**, 324–330 (2008).
57. Hernandez, D., François, P., Farinelli, L., Østerås, M. & Schrenzel, J. *De novo* bacterial genome sequencing: millions of very short reads assembled on a desktop computer. *Genome Res.* **18**, 802–809 (2008).
58. Butler, J. *et al.* ALLPATHS: *de novo* assembly of whole-genome shotgun microreads. *Genome Res.* **18**, 810–820 (2008).
59. Zerbino, D. R. & Birney, E. Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome Res.* **18**, 821–829 (2008).
60. Aury, J.-M. *et al.* High quality draft sequences for prokaryotic genomes using a mix of new sequencing technologies. *BMC Genomics* **9**, 603 (2008).
61. Reinhardt, J. A. *et al.* *De novo* assembly using low-coverage short read sequence data from the rice pathogen *Pseudomonas syringae* pv. *oryzae*. *Genome Res.* **19**, 294–305 (2009).
62. Schloss, J. A. How to get genomes at one ten-thousandth the cost. *Nature Biotech.* **26**, 1113–1115 (2008).
63. Altshuler, D., Daly, M. J. & Lander, E. S. Genetic mapping in human disease. *Science* **322**, 881–888 (2008).
64. Wang, L. & Weinshilboum, R. M. Pharmacogenomics: candidate gene identification, functional validation and mechanisms. *Hum. Mol. Genet.* **17**, R174–R179 (2008).
65. Haaland, W. C. *et al.* A–β– subtype of ketosis-prone diabetes is not predominantly a monogenic diabetic syndrome. *Diabetes Care* **32**, 873–877 (2009).
66. Tewhey, R. *et al.* Microdroplet-based PCR enrichment for large-scale targeted sequencing. *Nature Biotech.* **27**, 1025–1031 (2009).
67. Singh-Gasson, S. *et al.* Maskless fabrication of light-directed oligonucleotide microarrays using a digital micromirror array. *Nature Biotech.* **17**, 974–978 (1999).
68. Albert, T. J. *et al.* Direct selection of human genomic loci by microarray hybridization. *Nature Methods* **4**, 903–905 (2007).
69. Hodges, E. *et al.* Genome-wide *in situ* exon capture for selective resequencing. *Nature Genet.* **39**, 1522–1527 (2007).
70. Okou, D. T. *et al.* Microarray-based genomic selection for high-throughput resequencing. *Nature Methods* **4**, 907–909 (2007).
71. Porreca, G. J. *et al.* Multiplex amplification of large sets of human exons. *Nature Methods* **4**, 931–936 (2007).
72. Krishnakumar, S. *et al.* A comprehensive assay for targeted multiplex amplification of human DNA sequences. *Proc. Natl Acad. Sci. USA* **105**, 9296–9301 (2008).
73. Turner, E. H., Lee, C., Ng, S. B., Nickerson, D. A. & Shendure, J. Massively parallel exon capture and library-free resequencing across 16 genomes. *Nature Methods* **6**, 315–316 (2009).
74. Gnirke, A. *et al.* Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nature Biotech.* **27**, 182–189 (2009).
75. Olson, M. Enrichment of super-sized resequencing targets from the human genome. *Nature Methods* **4**, 891–892 (2007).
76. Garber, K. Fixing the front end. *Nature Biotech.* **26**, 1101–1104 (2008).
77. Petrosino, J. F., Highlander, S., Luna, R. A., Gibbs, R. A. & Versalovic, J. Metagenomic pyrosequencing and microbial identification. *Clin. Chem.* **55**, 856–866 (2009).
These authors describe the current state of metagenomics research and highlight the use of the Roche/454 platform for microbial identification through 16S ribosomal DNA phylogenetic analysis; other NGS platforms may be better suited for gene discovery efforts (see Table 2).
78. Lipson, D. *et al.* Quantification of the yeast transcriptome by single-molecule sequencing. *Nature Biotech.* **27**, 652–658 (2009).
79. Ozsolak, F. *et al.* Direct RNA sequencing. *Nature* **461**, 814–818 (2009).
80. Park, P. J. ChIP–seq: advantages and challenges of a maturing technology. *Nature Rev. Genet.* **10**, 669–680 (2009).
The article provides a comprehensive review of recent technological advances and challenges in genome-wide profiling of DNA-binding proteins, histone modifications and nucleosomes using NGS technologies (ChIP–seq).
81. Levy, S. *et al.* The diploid genome sequence of an individual human. *PLoS Biol.* **5**, e254 (2007).
82. Wheeler, D. A. *et al.* The complete genome of an individual by massively parallel DNA sequencing. *Nature* **452**, 872–876 (2008).
83. Iafrate, A. J. *et al.* Detection of large-scale variation in the human genome. *Nature Genet.* **36**, 949–951 (2004).
84. Sebat, J. *et al.* Large-scale copy number polymorphism in the human genome. *Science* **305**, 525–528 (2004).
85. Tuzun, E. *et al.* Fine-scale structural variation of the human genome. *Nature Genet.* **37**, 727–732 (2005).
86. Stranger, B. E. *et al.* Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science* **315**, 848–853 (2007).

©2010 Macmillan Publishers Limited. All rights reserved

87. Wang, J. *et al.* The diploid genome sequence of an Asian individual. *Nature* **456**, 60–65 (2008).
88. Kim, J. I. *et al.* A highly annotated whole-genome sequence of a Korean individual. *Nature* **460**, 1011–1015 (2009).
89. Ahn, S. M. *et al.* The first Korean genome sequence and analysis: full genome sequencing for a socio-ethnic group. *Genome Res.* **19**, 1622–1629 (2009).
90. McKernan, K. J. *et al.* Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. *Genome Res.* **19**, 1527–1541 (2009).
91. Pushkarev, D., Neff, N. F. & Quake, S. R. Single-molecule sequencing of an individual human genome. *Nature Biotech.* **27**, 847–850 (2009).
92. Mardis, E. R. *et al.* Recurring mutations found by sequencing an acute myeloid leukemia genome. *N. Engl. J. Med.* **361**, 1058–1066 (2009).
93. Lupski, J. R. *et al.* Complete genome sequencing identifies recessive alleles in SH3TC2 causing a CMT1 neuropathy. *N. Engl. J. Med.* (in the press).
94. Collins, F. S. & Barker, A. D. Mapping the cancer genome. *Sci. Am.* **296**, 50–57 (2007).
95. Drmanac, R. *et al.* Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science* 5 Nov 2009 (doi:10.1126/science.1181498).
96. Sanderson, K. Personal genomes: standard and pores. *Nature* **456**, 23–25 (2008).
97. Green, R. E. *et al.* Analysis of one million base pairs of Neanderthal DNA. *Nature* **444**, 330–336 (2006).
98. Briggs, A. W. *et al.* Targeted retrieval and analysis of five Neandertal mtDNA genomes. *Science* **325**, 318–321 (2009).
99. Ponting, C. P., Oliver, P. L. & Reik, W. Evolution and functions of long noncoding RNAs. *Cell* **136**, 629–641 (2009).
100. Carthew, R. W. & Sontheimer, E. J. Origins and mechanisms of miRNAs and siRNAs. *Cell* **136**, 642–655 (2009).
101. Barnes, C., Balasubramanian, S., Liu, X., Swerdlow, H. & Milton, J. Labelled nucleotides. US Patent 7,057,026 (2002).
102. Mitra, R. D., Shendure, J., Olejnik, J., Edyta-Krzymanska-Olejnik & Church, G. M. Fluorescent *in situ* sequencing on polymerase colonies. *Anal. Biochem.* **320**, 55–65 (2003).
103. Turcatti, G., Romieu, A., Fedurco, M. & Tairi, A.-P. A new class of cleavable fluorescent nucleotides: synthesis and optimization as reversible terminators for DNA sequencing by synthesis. *Nucleic Acids Res.* **36**, e25 (2008).
104. Yarbrough, L. R., Schlageck, J. G. & Baughman, M. Synthesis and properties of fluorescent nucleotide substrates for DNA-dependent RNA polymerases. *J. Biol. Chem.* **254**, 12069–12073 (1979).
105. Kumar, S. *et al.* Terminal phosphate labeled nucleotides: synthesis, applications, and linker effect on incorporation by DNA polymerases. *Nucleosides Nucleotides Nucleic Acids* **24**, 401–408 (2005).
106. McKernan, K., Blanchard, A., Kotler, L. & Costa, G. Reagents, methods, and libraries for bead-based sequencing. US Patent Application 11/345,979 (2005).
107. Macevicz, S. C. DNA sequencing by parallel oligonucleotide extensions. US Patent 5,969,119 (1995).
108. Mir, K. U., Qi, H., Salata, O. & Scozzafava, G. Sequencing by cyclic ligation and cleavage (CycLiC) directly on a microarray captured template. *Nucleic Acids Res.* **37**, e5 (2009).

Acknowledgements

I am extremely grateful to S.-M. Ahn, J. Edwards, J. W. Efcavitch, R. A. Gibbs, T. Harkin, E. Mardis, K. McKernan, D. Muzny, S. Turner and D. Wheeler for providing current performance data for the NGS platforms, and to the National Human Genome Research Institute for their support from grants R01 HG003573, R41 HG003072, R41 HG003265 and R21 HG002443.

Competing interests statement

The author declares competing financial interests: see Web version for details.

FURTHER INFORMATION

Michael L. Metzker's homepage: http://www.bcm.edu/genetics/?pmid=10947
1000 Genomes Project: http://www.1000genomes.org
Advances in Genome Biology and Technology meeting: http://www.agbt.org
The Cancer Genome Atlas: http://cancergenome.nih.gov
The Exome Project: http://www.nhlbi.nih.gov/resources/exome.htm
Human Microbiome Project: http://nihroadmap.nih.gov/hmp
National Human Genome Research Institute: http://www.genome.gov
National Human Genome Research Institute — Fruitfly Genome Sequencing: http://www.genome.gov/11008080
Nature Reviews Genetics poster 'Sequencing technologies — the next generation': http://www.nature.com/nrg/posters/sequencing/index.html
Personal Genome Project: http://www.personalgenomes.org

ALL LINKS ARE ACTIVE IN THE ONLINE PDF

©2010 Macmillan Publishers Limited. All rights reserved