nature biotechnology

The challenges of sequencing by synthesis

Carl W Fuller[1], Lyle R Middendorf[2], Steven A Benner[3], George M Church[4], Timothy Harris[5], Xiaohua Huang[6], Stevan B Jovanovich[7], John R Nelson[8], Jeffery A Schloss[9], David C Schwartz[10] & Dmitri V Vezenov[11]

DNA sequencing-by-synthesis (SBS) technology, using a polymerase or ligase enzyme as its core biochemistry, has already been incorporated in several second-generation DNA sequencing systems with significant performance. Notwithstanding the substantial success of these SBS platforms, challenges continue to limit the ability to reduce the cost of sequencing a human genome to $100,000 or less. Achieving dramatically reduced cost with enhanced throughput and quality will require the seamless integration of scientific and technological effort across disciplines within biochemistry, chemistry, physics and engineering. The challenges include sample preparation, surface chemistry, fluorescent labels, optimizing the enzyme-substrate system, optics, instrumentation, understanding tradeoffs of throughput versus accuracy, and read-length/phasing limitations. By framing these challenges in a manner accessible to a broad community of scientists and engineers, we hope to solicit input from the broader research community on means of accelerating the advancement of genome sequencing technology.

Attaining the Human Genome Project goal of sequencing the human genome and rapidly and publicly disseminating the data was a milestone in human biomedical research that was enabled by scientific, technical and cultural innovation. Central to the project's success was the development of robust, automated methods and technologies to identify the linear sequence of nucleotides. Recognizing the opportunities to use dramatically expanded sequencing technology in the subsequent phase of genomics research, in 2004 the National Human Genome Research Institute (NHGRI) of the National Institutes of Health (NIH) initiated a funding program with a goal of reducing the cost of genome sequencing to ~$1,000 in 10 years, with an intermediate goal of $100,000 by the end of 2009 (ref. 1; http://grants.nih.gov/grants/guide/rfa-files/RFA-HG-04-003.html). Numerous grant awards have been made in this program (http://www.genome.gov/10000368), which has stimulated a strong record of publications and patents and the successful commercialization of several second-generation sequencing platforms now in active use worldwide (e.g., https://www.roche-applied-science.com/sis/sequencing/index.jsp, http://www.illumina.com/pages.ilmn?ID=204, https://products.appliedbiosystems.com/ab/en/US/adirect/ab?cmd=catNavigate2&catID=604416, http://www.polonator.org/ and http://www.helicosbio.com/) with others in the wings (http://www.pacificbiosciences.com/, http://visigenbio.com/, http://www.intelligentbiosystems.com/ and http://www.completegenomicsinc.com/technology). At annual grantee meetings, open discussions of advances and challenges have stimulated collaboration and considerably accelerated research.

Several peer-reviewed articles have identified the strengths and limitations of commercial SBS platforms from a user's perspective[2–13] (for reviews, see refs. 5,14–16) and an assessment of the challenges facing nanopore sequencing has recently been published[17]. Here, we present current views of some of the investigators who received the technology development grants mentioned above on the underlying limitations and challenges of next-generation sequencing, with the goal of informing and engaging the broader research and engineering communities. Successful engagement requires the cross-pollination of ideas from experts across various disciplines to develop and identify solutions, as few scientists and engineers, including all but the most sophisticated users and active developers of technology, understand the full complexities of seamlessly integrating instrumentation, reagents and protocols necessary to promote scientific discovery.

SBS platforms

A common SBS strategy is to use DNA polymerase (**Fig. 1**) or ligase enzymes to extend many DNA strands in parallel. Nucleotides or short oligonucleotides are provided either one at a time or modified with identifying tags so that the base type of the incorporated nucleotide or oligonucleotide can be determined as extension proceeds.

SBS strategies may be categorized as either single molecule–based (involving the sequencing of a single molecule) or ensemble based (involving the sequencing of multiple indentical copies of a DNA molecule, typically amplified together on isolated surfaces or beads). They may be real-time (that is, with a free-running DNA polymerase given all nucleotides required) or synchronous-controlled (that is, using a priori temporal information to facilitate the identification process in a 'stop-and-go' iterative fashion). This can be achieved by using nucleotide substrates that are reversibly blocked or by simply adding only a single kind of nucleotide (e.g., dATP) at a time.

[1]GE Healthcare Life Sciences, Piscataway, New Jersey, USA. [2]LI-COR Biosciences, Lincoln, Nebraska, USA. [3]Foundation for Applied Molecular Evolution, Alachua, Florida, USA. [4]Harvard Medical School, Boston, Massachusetts, USA. [5]Howard Hughes Medical Institute, Janelia Farm Research Campus, Ashburn, Virginia, USA. [6]University of California, San Diego, La Jolla, Californi, USA. [7]Microchip Biotechnologies Inc., Dublin, California, USA. [8]General Electric Global Research Center, Niskayuna, New York, USA. [9]National Human Genome Research Institute, NIH, Bethesda, Maryland, USA. [10]University of Wisconsin-Madison, Madison, Wisconsin, USA. [11]Lehigh University, Bethlehem, Pennsylvania, USA. Correspondence should be addressed to C.W.F. (carl.fuller@alumni.upenn.edu).

Published online 6 November 2009; doi:10.1038/nbt.1585

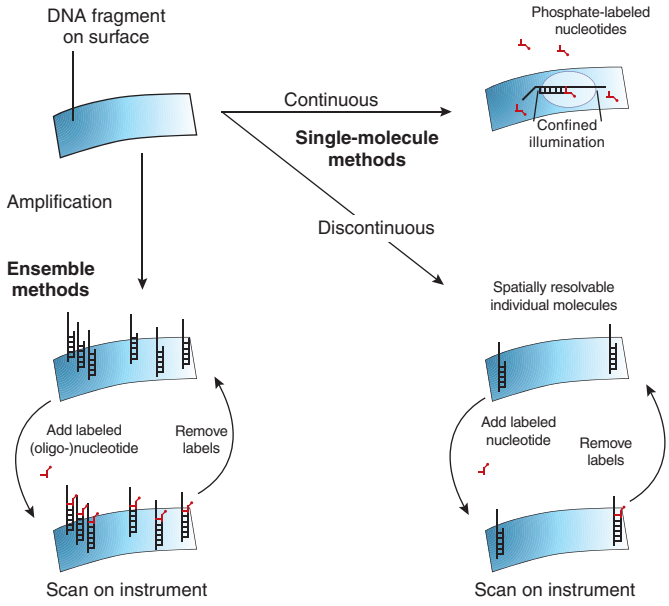


Figure 1 Schemes for SBS. Approaches include both single-molecule and ensemble methods requiring features populated with identical copies of DNA fragments. Ensemble methods monitor synthesis either using labeled substrates or detecting products of synthesis (pyrophosphate). Detection can be discontinuous (sequential) when labels can remain fixed for hours while a microscope scans up to hundreds of millions of sequence features[26,32]. It may also be continuous (simultaneous) as in the case where pyrophosphate, which can diffuse from the region of synthesis, is detected[23]. Similarly, single-molecule methods can be detected in a discontinuous (sequential) scanning mode[24] or continuous (simultaneous) mode by constant observation of synthesis of single molecules using labeled substrates[19,22]. The discontinuous scanning methods can take advantage of very large numbers of sequence features to assemble genome sequences from relatively short reads of 25–100 bases. The continuous-detection schemes read less than a million features but can achieve read-lengths of 3,000 or potentially more bases.

Real-time SBS requires identification of the newly incorporated nucleotide 'on-the-fly' without interrupting the synthesis process. The method requires the use of optical or physical confinement to distinguish free-solution tagged nucleotides from those bound and involved in synthesis[18–22]. **Table 1** shows the general approach used by several selected commercial platforms.

Synchronous-controlled approaches include pulsing the delivery of the nucleotide substrates (dNTPs) one at a time[23,24] or temporarily limiting extension using modified nucleotides, such as nucleotide reversible terminators (NRTs)[25–31], or cycling oligonucleotides[32,33] or omitting metal catalysts (Mg^{2+} or Mn^{2+})[34,35] during enzymatic labeling, followed by washing away excess reactants and imaging while synthesis is stopped. When only one of the four dNTP types is presented at a time, the mere fact that synthesis has taken place is sufficient to determine sequence. This can be done with relatively simple single-color fluorescence optics[24] or with coupled-enzyme chemiluminescence assays for pyrophosphate[23,36]. When all four types of dNTP are present at once, more complex, differential tagging schemes, usually with different fluorescence wavelengths to discriminate among the bases, must be used[19–22,27–32].

Challenges to SBS

Single molecule–based SBS and ensemble-based SBS face different problems. The former requires unique attention to such challenges as fluorescent labels, optics and instrumentation, whereas the latter requires further understanding of how phasing (that is, maintaining synchronous synthesis among all the identical templates of the ensemble) limits read-length.

These technical challenges have been accommodated to varying degrees in commercially available systems. The specifications of these systems, costs to use them and challenges to applying them to the production of particular biological data sets are under constant review[16] (**Box 1**). The success of companies in further enhancing the performance of their existing second-generation systems may well take place without the need to identify the challenges to the broader community of scientists and engineers. However, their ability to introduce third-generation sequencers may be predicated on how they address and solve many of the technical challenges through fundamental, innovative advances.

We detail below the technical challenges identified by NIH/NHGRI grantees as being associated with SBS. These include sample preparation, surface chemistry, fluorescent labels, optimizing the enzyme-substrate system, optics, instrumentation, understanding tradeoffs of throughput versus accuracy and read-length/phasing limitations (for additional challenges, such as data management and error correction, the reader is referred to recent reviews on sequence assemply[37] and mapping[38].

Sample preparation

With the development of relatively efficient enzyme, substrate and detection combinations, a potential cost bottleneck is the multistep sample preparation process. For ensemble-based SBS, sample preparation generally involves extraction and purification of DNA from tissues or other samples, fragmentation of the DNA, the repair of frayed ends in a polishing step, addition of adapters with ligases or transferases for solid-phase attachment and polymerase priming, and the clonal DNA amplification of a single DNA molecule to generate an ensemble containing thousands to millions of identical copies of the DNA molecule, bound to a surface in a localized packet (**Fig. 2**).

Several novel amplification techniques have been introduced, including the following: polony technology[33,39,40]; beads, emulsion, amplification magnetics (BEAM)[41]; emulsion polymerase chain reaction (emPCR)[23,42–44]; a cloning strategy developed for massively parallel signature sequencing (MPSS)[45]; and the PCR bridge amplification scheme[46,47]. The labor costs for these steps, which are borne by the user, have not been thoroughly addressed by system providers, especially if gel purification (as in paired-end tag libraries[33]) or emulsion formation are required. Efforts to automate as many steps as possible are underway. The amount of sample needed rises linearly with distance between paired-ends. Although pre-amplification has the disadvantages of being subject to contamination and the potential to introduce bias into a DNA library, its use has enabled successful sequencing with as little as 2 femtograms input (equivalent to a small bacterial genome molecule or 0.03% of a human diploid genome) with less than one amplification error per 100,000 base pair (bp)[48,49].

Rolling circle amplification (RCA) and multiple displacement amplification have also been used to prepare samples for SBS and other sequencing techniques[50–52]. To be efficient, these methods require displacement of the nontemplate strand. For circular templates smaller than ~100 nucleotides, displacement occurs spontaneously[53–55], but for larger DNA circles, it is preferable to use elevated temperatures or the DNA polymerase from bacteriophage Phi 29 (refs. 50,52,56–58). Phi 29 DNA polymerase has the ability to processively displace the nontemplate strand while replicating the template strand thousands of times. It also has proofreading activity and thus high fidelity. Overall signal strengths have also been improved by combining RCA with emPCR[51].

Ensemble-based SBS has the additional costs of library construction, single-molecule manipulation, amplification and associated technical

Table 1 Selected SBS platforms

Synthesis strategy	Company	Platform	Colors	Sequencing process	Amplification	Enzyme
Real-time	Pacific Biosciences	Zero-mode waveguide array	4	Continuous polymerization of labeled dNTPs	Single molecule	Polymerase
	Life Technologies; Visigen	Array of polymerase complexes	4	Continuous polymerization of labeled dNTPs	Single molecule	Polymerase
Synchronous-controlled	Life Technologies; ABI	SOLiD	4	Ligation of labeled 5 nt oligos	Yes, emulsion PCR	Ligase
	Danaher; Dover	Polonator	4	Ligation of fluorescently labeled 9 nt oligos	Yes, emulsion PCR	Ligase
	Illumina	Genome Analyzer	4	Polymerization using fluorescently labeled, reversibly terminating dNTPs	Yes, bridge amplification	Polymerase
	Roche; 454 Life Sciences	Genome Sequencer FLX	1	Polymerization of dNTPs added singly, luminescence detection of pyrophosphate	Yes, emulsion PCR	Polymerase
	Helicos	Heliscope	1	Polymerization of fluorescently labeled dNTPs added singly	Single molecule	Polymerase
Asynchronous with base-specific terminators	Various; requires high-resolution electrophoresis	Sanger dideoxy-sequencing	1, 4	Polymerization of fluorescently labeled ddNTPs with unlabeled dNTPs	Yes, clones or PCR	Polymerase

ddNTP, 2',3'-dideoxy-nucleotides.

problems (e.g., PCR amplification errors). Although extremely powerful, the multistep sample preparation and the clonal DNA amplification scheme are done manually and are very labor intensive. The complex workflow for ensemble-based SBS typically includes several enzymatic steps interspersed with cleanup steps using a spin column or beads. Mistakes at any step can ruin the preparation, and if not caught by the quality control checks, can waste not only the sample preparation reagents but also the downstream SBS reagents and instrument time. The sample preparation costs include several days of intense work by a well-trained researcher to prepare one to four libraries as well as the costs of kits (~$300 in reagents just for library construction and emPCR kits). In addition, the laboratory infrastructure requirements are significant, particularly when library construction is sensitive to contamination and necessitates work in a high-efficiency particulate air filtered hood or room and separate pre-amplification and post-amplification rooms.

A downside of ensemble-based SBS architectures is that sample preparation passes the analyte molecules through a single-molecule stage, only then to re-amplify them. The amplification from single molecules makes the process sensitive to amplification errors and products must be strictly isolated to reduce contamination of other libraries under construction. In addition to equipment routinely found in molecular biology laboratories, sample preparation involves quantitative PCR instrumentation, a fluorometer, a Bioanalyzer (Agilent Technologies) and a particle counter as well as other assorted supplies and reagents. Although challenging and still expensive, the process is successfully performed in research laboratories and improvements in methodology and automation are being addressed by both commercial suppliers and academic laboratories.

Single molecule–based SBS architectures that can acquire sequence information directly from individual DNA molecules without amplification[59] may reduce the sample preparation challenges, but introduce other problems for detection and resolution (see 'Optics challenges'). At first glance, the preparation of sample templates for single molecule–based SBS appears simple. But as with ensemble-based SBS, steps must be taken to provide DNA that is free from proteins or other contaminants. The collection of DNA fragments must provide full representation of the genome to be sequenced in fragments that both have usable ends, linkers or priming sites and are in the necessary size range. Furthermore, surfaces must be neither overpopulated nor underpopulated with template fragments. As with ensemble-based SBS, this requires careful control of the amount and size of DNA applied to the apparatus. Ideally, a surface could be devised that has properly spaced sites where single molecules of templates can be immobilized. Such a surface may increase the density of productive templates over a slide by two- to fourfold, and be immune to overloading (see also 'Surface chemistry challenges').

Widespread clinical applications will likely require streamlined, robust sample preparation techniques for supporting typical clinical loads. A 2-day sample preparation time by itself may restrict clinical applications and the ultrahigh throughput of gigabases of data per run must be matched to clinical needs, which may be for only thousands of bases of sequence per patient. One option would be an effective technology to judiciously add sequence tags to the fragments of several different samples, then combine and process the DNA samples (encoding and binning) together, separating the resulting sequences by their tags in the computer analysis phase[60]. This reduces the inefficiencies of ensemble-based SBS architectures (e.g., when abundant DNA is available to be sequenced as is the case when the patient is at hand). Any approach that directly sequences genomic DNA samples might offer a rapid and economical path to routine sequence-based clinical analysis. Additional challenges to overcome by sample preparation strategies include the following: first, handling double (or multiple) sequences due to diploidy and/or cross-hybridization of multigene family members; second, optimizing for the hybridization kinetics of single-copy genes; third, maintaining long-range haplotype information[48]; and four, generating binning reagents that give uniform yields (e.g., genome-tiling oligonucleotides).

Meeting the challenges of SBS sample preparation for research or clinical applications, regardless of whether the approach is ensemble-based or single molecule–based, will benefit from reducing costs and volumes of the biochemistry and automating the processes, thus eliminating the need for highly skilled molecular biology technicians. The throughput of the analysis portion must be matched by individual sample preparation instruments or banks of instruments. Applying massively parallel SBS methods to medical samples will require an affordable (e.g., $1,000) full genome or sequencing only specific regions of the genome (e.g., the 1% coding regions) from multiple patients together. Approaches to selecting

Box 1 How improvements impact cost

Because the technical challenges of SBS are not independent of one another, it is difficult to assign a metric that would identify how improvements for a particular challenge might affect overall system goals, such as increased throughput, lowered cost and reduced error rate. For example, polymerase incorporation rates of dNTP analogs and the bandwidth of the output amplifier on a CCD camera monitoring the real-time incorporation both affect throughput and one or the other may limit the sequencing rate[108]. Ultimately, information theory describes the overall relationship between throughput and accuracy and design teams are well advised to understand the theoretical limitations imposed by information theory and how throughput and accuracy create design trade-offs[84–103,106–108]. An example of this kind of trade-off is found in nature, which has optimized the coding for amino acids with nucleotide triplets[110].

Notwithstanding information theory, one can cautiously make some estimates of the impact of improvements in one challenge area by assuming 'ideal world' performance in the other challenge areas. Bifurcating between consumables (e.g., reagents and flow cells) and instrumentation can be a first-order approach to prioritizing the technical challenges with respect to cost. For example, ignoring consumable costs, a prediction of costs associated with the instrument (including purchase price, depreciation schedule, service contract pricing, performance-upgrade pricing and downtime) can be merged with various system throughput and accuracy targets to generate an estimated genome equivalent cost.

For the sake of illustration—without reflecting on the specifications of current commercially available next-generation sequencers—a $365,000 instrument with a depreciation lifetime of 5 years, with no service contract or upgrade costs, and with no downtime or bad samples, would cost $200 per day to run. If the throughput is 10 raw gigabases per day, and tenfold redundant sequencing is required for acceptable accuracy, then a 3 gigabase genome equivalent would cost $600 (not including consumable costs) and take 3 days. In contrast, a $500,000 instrument with a depreciation lifetime of 5 years and a $10,000 per year service contract but with free upgrades, 10% downtime, 3 gigabase per day throughput and 30× redundant sequencing, a 3 gigabase genome equivalent would cost $10,000 (not including consumable costs) and take 33 days. Through 'crystal-ball' spreadsheet analysis, one can project how instrument-related costs such as manufacturing costs (parts, labor and overhead) or the instrument robustness (which contributes to downtime costs as well as service contract costs), can directly affect the genome equivalent cost as well as the sequencing throughput.

appropriate subsets of genome sequence by hybridization capture or targeted circularization have already appeared[61,62]. Additional challenges include extending the length of clonally amplified DNA, minimizing sample preparation reagent and equipment costs, preventing cross-contamination, and barcoding[63] when using multiple patients' samples. Library and amplification reactions, whether done in multiwell plates or flow cells, typically occur in small microliter volumes and thus reagent costs should be similar to conventional molecular biology reagent costs (e.g., $1 per reaction step). Opportunities exist for using single cells and even '*in situ* tissues' as RNA or DNA sources.

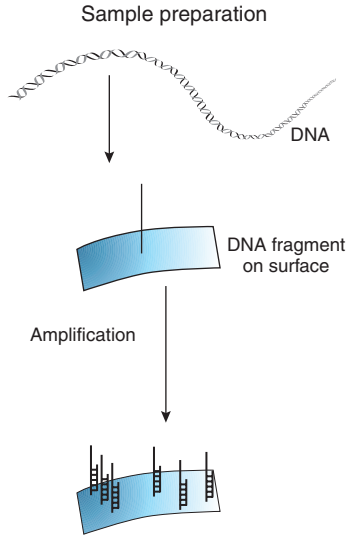


Figure 2 Sample preparation includes arraying individual fragments of DNA on beads or other solid surface. For ensemble methods the fragments are amplified, providing a collection of identical copies for sequencing.

Surface chemistry

Surface chemistry strategies arise from the critical interface between the sequencing biochemistry and detection. Trade-offs or compromises between the ideal characteristics of these two design elements require close communication and understanding among the molecular biologists, the chemists and the optical and mechanical engineers of the design team. All SBS schemes involve the use of surface-bound components that provide a means for parallel synthesis of DNA molecules, either singly or as ensembles of identical sequences (**Fig. 3**). The surfaces also provide a structure for optimizing imaging, and for flowing-in substrates and removing products. Surfaces used include flat slides, beads and specialized fabricated structures (e.g., waveguides and microfluidic channels)[18,19].

The challenges of surface chemistry include providing surfaces compatible with enzymatic processing of nucleotides along with the DNA, eliminating stray 'sticking' of dye molecules and maximizing the density of SBS features over the surface. For this reason, structures are typically coated with a hydrophilic, functional surface layer, like those used for microarrays, designed to tightly or covalently bind the molecules of interest (to minimize loss of molecules and thus of signal) but to be inert to binding other materials during sequencing. These aspects of surface chemistry are even more important in single-molecule forms of SBS[18,19,64]. To immobilize single polymerase enzyme molecules in the bottom of zero mode waveguides, a surface coating was developed that sticks to the metal sides of the waveguide and repels proteins but doesn't coat the glass floor of the waveguide chamber[19]. In addition, the topology of surfaces for holding the growing chains of SBS experiments can influence the density of chains distributed within the imaging area of the instrument, directly influencing imaging efficiency.

The surfaces of slides or specialized structures are commonly treated with silanes, polyethylene glycol, pluoronics, proteins or other surface-modifying agents to produce highly uniform, biocompatible surfaces[21,46,59] because they must both anchor biomolecules and ensure enzymatic activities. These chemistries inherently do not have

well-defined stoichiometry and attaining reproducibility may require substantial optimization whenever any change is made in the surface chemistry, the biochemistry or the detection conditions.

Another challenge in managing surfaces for both single molecule–based and ensemble-based SBS relates to the effective use of geometry, which in turn relates to the cost and speed of the imaging system. Primers, target DNA or polymerase are distributed, often at random, across two dimensions to a surface, with Poissonian statistics driving the throughput trade-off between lower-density populations of DNA-polymerase complexes versus spatial aliasing of overpopulated surfaces. This often results in densities less than a third of that which is maximally resolvable. Resolving this challenge requires surface chemistry strategies that enable ordered, closely packed arrays with better density than random attachment or directed nonrandom placement of DNA-polymerase complexes[19].

Resolving the challenges specific to single molecule–based SBS requires surface chemistry that eliminates stray sticking of dyes, enzymes or DNA and improved ways of obtaining high-density distributions of single molecules of target DNA. For ensemble-based SBS methods, the accumulated experience with microarray technology makes it easier to build useful arrays of DNA molecules suitable for manipulation and detection. Small beads[32–34,45,65,66] provide one convenient method for the rapid assembly of high-density arrays to provide the necessary solution to interfacing the sequencing biochemistry and the detection. In methods involving both beads and slides, the surface chemistry of the beads needs to be optimized to prevent nonspecific binding between the relatively large bead and the slide surface, yet allow the bead to approach the surface to bind with DNA[22].

Several of the above challenges arise because the molecular presentation approaches work hand in hand with detection schemes that are currently bound by resolution limits of light microscopy. The trade-off between biochemistry and detection may be addressed by more efficient SBS systems that use radically different detection schemes such as electronic read-out techniques that are engendered within the same sequencing device.

Fluorescent labels

The commercial introduction of several ensemble-based SBS systems makes it clear that the fluorescent dyes used for electrophoretic sequencing and microarrays are well suited for ensemble-based SBS. With much more limited commercial experience with single molecule–based SBS systems, there still may be a need for careful choice of dye tags when every relevant single molecule needs to be detected. Failure to detect a single dye molecule may result in a deletion error (false negative) and detection of a stray molecule may similarly result in an insertion error (false positive).

Single molecule–based SBS approaches require the sensitive and versatile detection offered by fluorescent labels. Some schemes require four distinct fluorescent labels[26,32,50], each detected at a different wavelength. Others get by with a single fluorescent tag[23,24] or rely on additional dyes used in energy-transfer schemes[2]. In all cases, the choice of dye is largely dictated by sensitivity of the optical system, in particular the detector. Because detection can involve complex design trade-offs with respect to dye characteristics, it is important to identify suitable dyes as early as possible in the design process of single molecule–based SBS.

With fluorescence, suitability of a particular dye is a matter of at least a half-dozen characteristics beyond simple quantum efficiency. These characteristics include the ability to capture excitation photons (of the wavelength available in the instrument), rapid release of emitted photons, quantum efficiency and photostability. The efficiency of capturing photons of the excitation wavelength is conventionally expressed as the extinction coefficient or molar absorptivity, and varies with wavelength. For example, the molar absorptivity of fluorescein at 488 nm is about 80,000 $M^{-1}cm^{-1}$. This value is sufficient for sensitive measurements and there are several families of dyes (e.g., the cyanine dyes) with two- to threefold greater molar absorptivity[67].

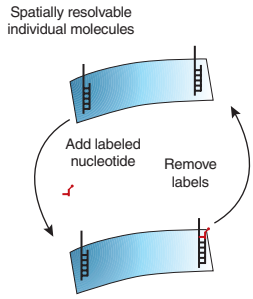


Figure 3 Surface chemistry. Surfaces may assist in producing optically resolvable sites for sequencing and must be compatible with the enzymes used for synthesis. Nonspecific binding of labels must also be minimized.

Once a photon is captured by a fluorescent dye, it remains in the excited state for a characteristic period of time (the fluorescence lifetime of the dye), which can range from nanoseconds to milliseconds. For applications like single molecule–based sequencing, where data collection must be completed quickly and a large number of photons must be collected to distinguish the different dyes, lifetimes of less than about 5 ns are preferred. With adequate illumination, a single dye molecule can cycle between the ground and excited states millions of times per second, providing a sufficient number of emitted photons for reliable detection within a few milliseconds.

The photostability of a dye, particularly when performing single-molecule detection, is more difficult to predict based on its structure. Because the dye molecules will reside for a substantial fraction of the time in the excited state, when a high photon flux is needed, the stability of the excited state becomes critical. Some dyes will cross over from the excited state to a long-lived triplet state, during which no photons are emitted. Other dyes may lose activity by isomerizing while in the excited state. When these two processes are reversible, single molecules of dye can appear to 'blink' off and on during extended observation periods and may result in sequence errors. Therefore, for single molecule–based SBS, where every fluorophore counts, fluorophore purity and storage conditions are also crucial. 'Dead' or bleached fluorophores (duds) and blinking are not an issue with ensemble-based SBS; however, for single molecule–based SBS, bleaching and blinking would cause false negatives unless multiple fluorophores can be used[43]. Fortunately, the frequency of crossing over to triplet states is usually very low and isomerization can be avoided by designing dyes with rigid structures.

For single molecule–based SBS, contamination of labeled dNTPs by unlabeled dNTPs (e.g., impurities or hydrolysis products) is another potential source of false negatives. For phosphate-labeled dNTP substrates, a phosphatase can be used before or during the sequencing reaction to convert background unlabeled dNTPs to nucleosides (which are not substrates for polymerases) while not affecting the desired, phosphate-labeled dNTPs[21,68].

Perhaps the greatest challenge is the elimination of stray signals from dye molecules that stick to the sequencing surfaces. Efforts to make surfaces more uniform and nonattractive to dyes may be required. In addition, the dye parameters (e.g., charge or hydrophobicity) need to be carefully chosen during dye design to avoid ionic attraction or promote

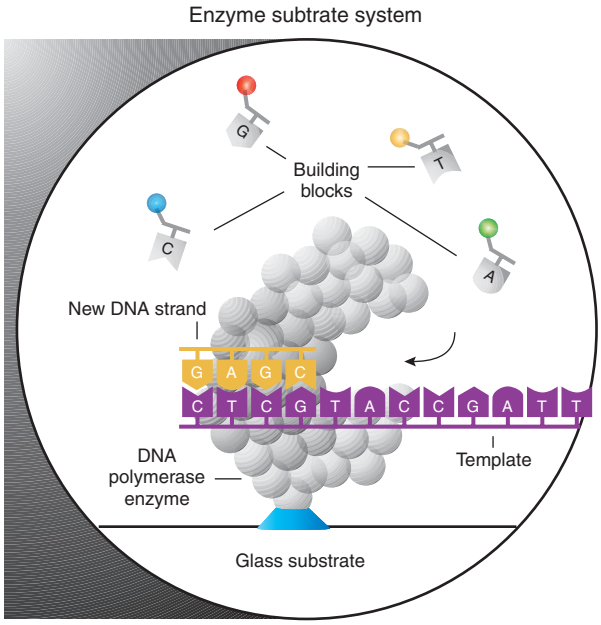


Figure 4 Enzyme substrate system. Although the enzymology of DNA synthesis is well understood, modified building blocks must be used for labeling and controlling the rate of synthesis. For example, labels have been attached to nucleotides at the phosphates, the sugars and the bases to get suitable activity. This is illustrated here for a single-molecule real-time SBS method. The polymerase is modified by attachment to substrate, and the nucleotides have fluorescent tags attached to the phosphates. Reprinted with permission from Pacific Biosciences.

repulsion from as many surfaces as possible without interfering with enzyme functionality.

The enzyme-substrate system

SBS is based on the stepwise enzymatic synthesis of DNA (or potentially RNA) complementary to the template DNA to be sequenced (**Fig. 4**). For a subset of polymerase-based SBS approaches, the identity of a template base is determined by observing the stepwise addition to the synthesized DNA strand of a labeled dNTP substrate complementary to that template base (pyrosequencing SBS approaches use a polymerase in conjunction with unlabeled dNTP substrates with the template base identified by quantitatively assaying pyrophosphate release). When using a polymerase with a labeled dNTP substrate system, the label needs to be removed or deactivated (e.g., photobleached) once it is detected, so that the next stepwise addition can be observed without background. This requirement for efficient removal has led to the development of schemes in which immobilized templates or enzymes are used, with substrates and products flushed in and out of the system, in a process synchronously coordinated with detection. To obtain long read-lengths using ensemble-based SBS, repeated additions must be accommodated with virtually 100% stepwise yields so that chain extension is synchronous over all the molecules of the sample. This includes both the enzymatic addition and any subsequent chemical, enzymatic or photolytic steps that may be needed to unblock the substrate or remove the dye for the next addition.

This need for rapid, nearly quantitative activity means that sequence determination depends on the characteristics of the polymerase enzyme and the tagged and often blocked (e.g., by 3′ modification) dNTP substrates used for chain extension. The relative activity between the polymerase and the dNTP substrates depends on the nucleotide base, the tags and linker, the template sequence context, the presence of competing nucleotides or reaction products and reaction conditions. It may also be affected by the practice of adding a single nucleotide rather than a mixture of all four. This may increase the incidence of misincorporation, and additional analysis may be needed to properly interpret the data[22,23]. Optimizing a sequencing scheme may require exploration of the kinetics of several polymerases under a range of conditions for each tagged nucleotide and with a range of template sequences. Nucleotides and polymerases that appear to work well in isolation may not do so when used in combination (C.W.F., unpublished data).

With a nearly infinite variety of reaction conditions, dyes, linkers and blocking groups potentially available for the four dNTP substrates, one must choose a limited, compromise subset for testing new enzymes and other system components. Although full optimization in such a large experimental space may not be possible, a variety of modified nucleotides have been used successfully for DNA sequencing with polymerases including ones labeled or modified on the base[24–30,68–70], sugar[25–29,71–74] and phosphate[34,64,75–80]. A few principles can be used as guides to optimization, but in general polymerases are remarkably tolerant to modifications of the substrate nucleotides. For example, attaching a dye through linker arms to the major-groove side of the nucleoside base usually affects activity with polymerase by less than one order of magnitude, depending somewhat on the net electrostatic charge of the dye and linker. Modifications on the phosphate are also remarkably well tolerated and several sequencing schemes have been developed to take advantage of this architecture[19,22,68]. Thus, it is reasonable to collect a number of polymerases and modified nucleotides and test them under a variety of reaction conditions, paying particular attention to reaction rates and yields.

Whereas it may be more challenging to obtain novel polymerases than to synthesize labeled nucleotides, many different DNA polymerases have already been isolated and studied, and they are available for experimentation. Because they are useful for sequencing, amplification and cloning, many diverse polymerases are available commercially as well. There are even some generalized observations that can facilitate optimization. First, most polymerases have exonuclease activity (exo$^+$). Although these exo$^+$ polymerases are essential for repair and high-fidelity DNA replication, the exonuclease activity is usually detrimental to sequencing. Fortunately, the active site for nuclease catalysis is typically distinct from the polymerase site and can be readily recognizable in primary sequence. Inactivation of the exo$^+$ site can then usually be achieved by simple, directed mutagenesis or deletion. Many commercially available DNA polymerases and reverse transcriptases have reduced exonuclease activity (exo$^-$); these should be among the first ones tested with any modified nucleotide.

Several polymerase-based SBS schemes (e.g., synchronous-controlled SBS) require using blocking groups that allow the addition of a single nucleotide at a time. Of particular interest in these schemes is the deoxyribose sugar because it is part of the backbone chain of DNA. A hydroxyl group at the 3′ position of the sugar is required for further extension of the nascent chain so this position is ideal for blocking extension after each addition. The activity of most DNA polymerases is reduced by more than 10,000-fold simply by replacing the 3′-hydroxyl of dNTP with a hydrogen (resulting in a 2′,3′-dideoxy-nucleotide). Similar responses are observed when the 3′-hydroxyl group is replaced by larger groups or when the 2′ position is altered. One exception to this general rule is the well-known activity of T7 DNA polymerase with 2′,3′-dideoxy-nucleotides. This polymerase has the hydroxyl of a tyrosine in the binding site of the nucleotide[73,74] instead of the consensus phenylalanine. This change compensates for the lack of an oxygen atom at the nucleotide 3′ position.

Even with relatively simple activity assays, the process of finding the critical tyrosine amino acid and verifying its importance for T7 DNA polymerase was a difficult and time-consuming task. Although this single amino acid change provides a very useful benefit, it is much more likely that the selection of polymerases with improved functionality with other modified nucleotides will require multiple amino acid changes and result in more gradual improvements in activity. Strategies for creating libraries of mutated polymerases with improved properties for using other modified nucleotides may require a combination of both random and directed steps[81]. In addition, a survey of reaction conditions involving dNTP concentration, metal choice (Mg^{2+} or Mn^{2+}), metal concentration, pH, operating temperature and salts (particularly ammonium and sulfate) should be done so that beneficial changes are not overlooked. Fidelity and lack of terminal transferase functionality are also critical to successful polymerase selection. Polymerases with faster catalytic rates could increase both the speed and the phasing efficiency of ensemble-based SBS reactions by increasing the percentage of templates going to completion on each cycle.

Ligase-based SBS architectures pose similar optimization issues, but because an oligonucleotide can be tagged far from the 5′ phosphate position where the reaction takes place, specific interactions between tags, blocking groups and ligase may not arise. Fewer DNA ligases are available on the commercial market, but some from bacteria, archaea and bacteriophage are available for testing. Because SBS that involve ligation use a diverse mix of 6-mers to 9-mers, efforts to get the enzyme substrate pairs to be more uniform in catalysis with respect to melting temperature (or GC content) could reduce errors or increase speed[33,63]. As with polymerization, the most important criterion is to maintain near-quantitative yields so that all the copies of a template in a feature remain in phase so that read-length and accuracy are not compromised.

Optics

The optical detection limits for real-time and synchronous-controlled SBS sequencing differ as do methods based on single-molecule detection compared with ensemble-based approaches. To date, single-molecule methods require high (>1.3) numerical aperture (NA) microscope objectives for detection to compensate for the limited emission and high speed. The maximum useful field of view is ~2,000 pixels diameter. A four megapixel camera captures all available image area and the camera read rate determines the ultimate throughput. For the preferred commercially available electron multiplying charge-coupled device (CCD) cameras, current limits are 1 megapixel size and 35 million pixels per second read rates. One bottleneck related to the read rate is the bandwidth of the output amplifier on the CCD camera. Synchronous-controlled methods rely on a moving stage to sequentially acquire many thousand fields of view, so stage move and stop times also become critical. For synchronous, single-molecule methods, the objective field of view limit remains important. For synchronous, ensemble-based methods, collection optics with lower NA and much larger fields of view can be used. For example, a commercial f#/1 (NA = 0.5) 4 × 5 camera lens has a field of view >40,000 pixels wide, far beyond any single element camera chip. Going forward, improvements in molecule emission (photons per second and photobleaching limits), camera size and read rate, collection optics efficiency and field of view, and stage move times will all affect sequencing rate in the complex parameter space of sequencing performance.

Throughput for all the above methods is improved if fewer pixels per sequencing site can be used. Random arrays 'consume' ~100 pixels per sequencing site. New CCD camera technology that would allow the simultaneous detection of one million sequencing sites would need to have ~20 million pixels for zero-mode waveguide technology. This assumes approaching the theoretical limit of random distribution in which 37% of waveguides are loaded with exactly one enzyme molecule[19]. With array assembly methods enabled by micro- and nanofabrication technology, it may be feasible to array the DNA samples such that the spacing and geometry of the individual DNA templates match the resolution of the imaging lens, excitation pattern and the camera to dramatically increase the imaging and data acquisition efficiency[62,65]. It is currently feasible to image each DNA template with only four pixels (X.H., unpublished data). Further imaging improvements may enable detection of each template with only one pixel. However, when using fluorescently labeled dNTPs with four different wavelength spectra, reducing chromatic aberration in the optics to achieve this resolution may introduce considerable cost, depending on the wavelength separation. Substantially overlapping fluorescence spectra and the potential of chromatic aberration may compromise read accuracy. New far-field optics, such as synthetic aperture optics[82,83] that enable optical scanning of a very large area of arrays without the use of mechanical stages may also enable rapid excitation for sequence readout.

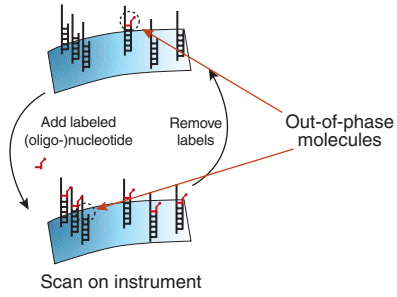


Figure 5 Read-length and phasing. For ensemble sequencing methods, thousands of identical copies of a DNA sequence are synthesized together. Whenever synthesis lags behind or steps ahead, signal is lost and background increases. Ultimately, this limits the length over which synthesis can be read.

Successfully combining an increase in the detection area containing thousands of sequencing sites along with an increase in the percent usable chip area may promise to increase SBS throughput while maintaining data quality. But the engineering required is not trivial and is ultimately subject to limits imposed by information theory, as the task of fluorescence systems is to transduce the stochastic (uninformative) photons emitted by a light source into ordered (informative) photons. Information theory imparts, for a given set of engineering design parameters and fluorescent dye characteristics, a relationship among input power, sampling bandwidth, throughput and accuracy. Certainly increasing the input power should increase the signal-to-noise ratio, and thus the overall throughput for a given accuracy. This can be done by merely adding more sequencing instruments (which adds cost proportionally) or by increasing the laser power in a given sequencing instrument, which will increase the rate of photobleaching. Rapid photobleaching can be mitigated by the use of multiple fluorescent dyes (possibly attached to nanoparticles) for the critical stepwise detection events, particularly for sequencing single DNA templates.

For example, information theory predicts that a single molecule–based SBS system designed for extremely long reads, 100 milliwatts of total input laser power split into four signal channels (A, C, G or T), each of which irradiate a large number of sequencing sites (within a 100 μm diameter spot size), using state-of-the-art collection optics, a high-sensitivity CCD camera with 20 ms integration and highly efficient fluorophores can yield ~30 bases per second per sequencing site with an error rate <0.28% (L.R.M., unpublished data; refs. 29–36,84–103).

However, increasing the integration period to 50 ms 'rate limits' the throughput to 20 bases per second but decreases the error rate by over five orders of magnitude by reducing the overlap between signal and background. Reducing laser power to 50 milliwatts also has a complex effect, increasing error rates by over four orders of magnitude but reducing throughput only by about half.

A similar analysis can be performed for systems that have short reads and so require a moving CCD camera stage to generate high throughput. The optimum design parameters for various kinds of systems will differ, but the relationship among input power, sampling bandwidth, throughput and accuracy and other parameters is still ultimately dictated by information theory.

Currently, adequate signal-to-noise characteristics exist to detect single molecules directly given enough sampling time, so one might expect that detection in ensemble-based SBS approaches (e.g., multimolecule polonies) would have little room to improve. However, extra signal (or greater signal-to-noise) could permit use of faster and/or cheaper CCD or complementary metal-oxide semiconductor (CMOS) detectors. The trade-offs among pixel size, pixel-to-template ratio, array size, readout speed, spectral responsivity and readout noise may be affected substantially by the signal level generated by the fluorescent tags. In the case of single molecule–based SBS, the throughput is particularly limited by the number of photons per unit time that one can detect from single small organic fluorophore molecules. Increasing the speed of incorporation from 10 bases per second to 50 bases per second would require some other kind of detection technology.

Notwithstanding the ability to detect single molecules directly, the necessity for detecting and identifying single molecules for single molecule–based SBS brings additional challenges[19,104]. These challenges include scanning large areas with sufficient sensitivity to detect, and in some cases spectrally identify, single dye molecules to distinguish real from stray molecules (e.g., collecting fluorescence from smaller and smaller optical excitation volumes to reduce the background signal associated with the large fraction of labeled nucleotides required to provide adequate enzyme kinetics) and to complete the scans quickly enough to avoid damage to the dyes. With single molecule–based SBS schemes, it is difficult to determine when an error might have been encountered in the data stream. Dark periods with no signal due to the blinking of the dye molecules are normal events, often (but not always) safely ignored. One simply waits until a signal is detected, and then interprets it appropriately.

Overcoming the challenges of single molecule–based SBS will likely require a combination of careful analysis of the fluorescence properties of dyes that are used, careful control of the environment used for observation, high-efficiency optics, reliable methods to distinguish dyes of interest from those free in solution and measures to verify image analysis results. In some cases, it may be possible to either resequence the same template (e.g., using RCA sequencing with strand-displacement on the same template[50]) or substantially extend read-length to reduce errors (e.g., false negatives due to missed nucleotides).

One might also envision using a multiply labeled spectrally encoded bead or nanoparticle in conjunction with single DNA template sequencing to retain the signal-to-noise strengths of ensemble-based detection. With respect to multiply labeled beads, the current state-of-the-art detection is well below (50-fold) the maximum signal achievable by packing conventional fluorophores into a size near the diffraction limit (G.M.C., unpublished data). One could attempt to approach this maximum signal by increasing the packing density limit while avoiding fluorescence quenching and/or steric interference with enzyme functionality.

Instrumentation

Design of sensitive, high-speed instruments like those to be used for SBS is always a complex process. A system for SBS involves producing suitable template strands for sequencing, sometimes amplifying them, and finally delivering them to the active portion of the instrument. Synthesis reagents, such as enzymes and/or tagged nucleotides, are subsequently introduced. Detection is done continuously or, more commonly, in cycles after removal of excess tagged materials. Finally, data are collected from as many template sequences as possible in parallel, often requiring rapid determination of color and intensity of signals. Because sequence is usually built by assembly of sequential imaging steps, images must be brought into register and individual templates identified repeatedly.

The challenges are to introduce DNA, manipulated in microgram quantities and in microliter volumes, into sensing volumes having just one to a few thousand molecules and ~10^6-fold smaller than the initial sample volumes. These challenges have been overcome by using a variety of manipulations, most notably PCR, on surfaces and on beads. Separation of bound, reacting molecules from those free in solution has also been achieved by using optical methods (e.g., zero-mode waveguides[18,19]) and immobilization on beads. These methods are not simple to monitor and evaluate, and one can expect that further progress in the area of manipulating templates, substrates and enzymes within detection instruments will be required to achieve efficient, trouble-free sequencing experiments whose costs readily match clinical expectations.

Several technologies need to be integrated into the instrumentation required by SBS, including electronics, optics, mechanical (static and dynamic design), packaging, embedded software and graphical user interface software as well as various technologies specific to the sample and reagent presentation such as micro- or nanofluidics, nanotechnology, pumping and valving[105]. The amortization schedule of instrumentation cost is affected by component reliability, lifetime, availability (e.g., second sourcing) as well as system design reliability. In some cost models, the ultimate cost bottleneck may not be dictated by the sample preparation or the reagents, but rather by the tradeoffs among sequencing speed, instrument reliability (servicing costs and/or amortization lifetime), and initial instrumentation cost. SBS instrumentation may not be able to achieve the economies of scale associated with other complex systems that can be produced in high volume (e.g., computers or automobiles) to significantly drive down component and manufacturing costs.

Throughput versus accuracy

All designers and users of complex sequencing systems strive to increase system throughput while enhancing, or at least without compromising, accuracy. When system efficiency approaches the limits dictated by information theory, all attempts to improve throughput will result in reduced accuracy. For this reason, the multidisciplinary teams should build models based on information theory to test system efficiency, and heed the advice of user-informaticians[84,106–108].

The concepts of sequencing throughput, read-length, coverage or even cost are meaningless without a definition of accuracy. The 'standard' proposed in the NHGRI grant solicitations included both per-base sequence accuracy and assembly statistics[108]. When the first solicitation was published in 2004, achieving that quality required some effort with Sanger/capillary array methods, and it has not been met for *de novo* sequencing of a human-sized genome by any of the current next-generation sequencing technologies. Furthermore, that standard is insufficient for most medical studies. The Archon X prize attempts to specify a goal for an affordable genome that might approach utility for specific disease-gene studies or possibly for individual patient diagnostics (http://genomics.xprize.org/) using the following criteria: (i) 98% of the genome covered; (ii) 10 days per 100 human diploid genomes at

$10,000 each; (iii) "no more than one error in every 100,000 base pairs" (note that this differs from "no more than one error per 100,000 averaged over the whole genome"); and (iv) "a rearrangement or haplotype error counts as one error; insertion and deletion errors (indels) count as the sum of each base in the indel."

To properly assess throughput versus accuracy versus cost, a comprehensive cost model needs to be developed that includes accuracy as a variable parameter. Such a model would both be anchored by actual run costs on a system where actual reagent costs are relatively transparent (both including and excluding marketing and royalty costs) and capture the network of cost inter-dependencies among components. It would also permit the determination of a variety of cost projections with different parameters, enabling comparisons of trade-offs (speed, percent and which parts of the genome covered), analysis of rates of various types of errors (e.g., point, indel, rearrangement or haplotype phase errors) and assessment of how changes in sequencing system parameters affect cost and accuracy. A particularly important error to consider is in the discrimination of heterozygotes from homozygotes. Typically, heterozygotes are undercalled, and incorrect calls of homozygous for a deleterious recessive allele cause a false-positive indication, whereas incorrect calls of the nondeleterious allele would result in a false-negative diagnosis. For single nucleotide polymorphism (SNP) genotypes, this problem is fairly independent of read-length, enzymology or single molecule–based SBS versus ensemble-based SBS. Calling accuracy is highly dependent on redundancy (at least 30-fold coverage of each region or 15-fold coverage of each allele is desirable to achieve Archon X prize accuracy levels) and somewhat dependent on raw error rate because sequencing errors can be seen two (or more) times at a given base and thus misinterpreted.

Signal-to-noise ratios have an impact on the theoretical throughput of information for a given accuracy level and, as earlier emphasized, an enhanced understanding of this relationship using information theory, particularly for single molecule–based SBS embodiments, would be instructive in providing a sense of whether throughput can be improved for a certain level of accuracy or whether one is already near the limit[106]. Information theory can also assess the impact of redundant sequencing coverage to increase accuracy. Stochastic errors associated with SBS techniques may be eliminated by means of minor redundant (low-coverage) sequencing to generate high-accuracy finished sequence as contrasted to systematic errors. For those synchronous-controlled SBS systems that have excess signal-to-noise in their early cycles, one may be able to enhance overall throughput without sacrificing accuracy by reducing the scan time for the early cycles.

Read-length and phasing limitations

Ensemble-based SBS includes methods of creating collections of identical sequences (by PCR, RCA or other processes) and determining their sequence by synthesis of the complement in a stepwise, synchronous fashion. The result is an 'average' sequence signal from all the copies present, and typically accuracy drops with successive steps as synthesis on some templates lags behind that on other templates (**Fig. 5**). This trend to lose synchrony can establish the limit of accurate read-lengths. Single-molecule or real-time SBS methods have entirely different factors determining read-length and accuracy. In this case, read-length depends more on overall incorporation-cycle efficiency and reliability of accurate detection of the dye tags. Termination of chain growth on a single template molecule ends a read in the single-molecule case, whereas for ensemble sequencing it can merely reduce the signal level.

The ensemble pyrosequencing method used by 454/Roche (Basel) has been shown to maintain reasonable phasing through 400 cycles of addition using natural dNTP substrates[23]. It is unknown whether this is an upper limit or whether a similar number of cycles could be achieved for polymerase approaches using labeled or blocked dNTPs or for ligation reactions with 5-mer or 6-mer additions, either of which theoretically could mean five- to sixfold longer reads.

Long read-lengths require an effective solution to the synchrony problems in ensemble-based SBS. All primers must be extended by one nucleotide. Extension must then completely stop for a sufficient amount of time to collect the data needed to call the added nucleotide. Then, whatever chemistry is required to permit extension to resume must have a 100% yield. Such problems occur in analogous forms in other settings (e.g., the solid-phase synthesis of DNA or the sequential Edman degradation of polypeptides), and one expects that they will be solved in ensemble-based SBS in the same way that they have been solved in other settings—by innovation of new chemical and enzymatic reagents and system integration.

A better understanding of exactly which parts of the sequencing process can be asynchronous is critical. As an example, in four-color DNA SBS using cleavable fluorescent NRTs, to negate any lagging fluorescence signal that is caused by a previously unextended priming strand, a synchronization step can be added to reduce the amount of unextended priming strands after the extension with the fluorescent nucleotides. For example, Ju *et al.*[29] have described a protocol in which a synchronization reaction mixture consisting of four 3′-*O*-modified-dNTPs, which have a higher polymerase incorporation efficiency due to the lack of a fluorophore, is used along with the DNA polymerase to extend any remaining priming strand that has a free 3′-OH group. The extension by 3′-*O*-modified-dNTPs also enhances the enzymatic incorporation of the next nucleotide analog because after removal of the 3′-*O*-capping group, the DNA product extended by 3′-*O*-modified-dNTPs lacks a modification group.

Conclusions

The challenges identified above have both near-term and long-term implications and trade-offs—depending on the particular SBS strategy (real-time or synchronous-controlled), the particular stepwise approach (if synchronously controlled), the stepwise manner in which dNTPs are delivered (simultaneously or sequentially in time) and whether the approach is ensemble-based or single molecule–based. Building and integrating instrumentation around these biochemical systems further exposes the weakest links in the individual challenge areas.

Several next-generation SBS approaches have already been commercialized. These have dramatically increased throughput and have substantially reduced the cost of sequencing compared to traditional Sanger-based methods. For the near-term, these commercial systems are constantly undergoing further improvements by the manufacturer as well as their customers.

When NHGRI launched the current technology development program, it linked the cost goals to a quality metric; the technologies should be capable of producing genome sequence at least as good as the mouse draft genome sequence assembly published in 2002 (ref. 109). Through implementation of SBS it appears likely that the goal of achieving human sequence of that quality for about $100,000 will be achieved imminently; precise costs are difficult to validate for technology that is changing so fast. Even so, at the current state of the technology, that success is highly dependent upon complex and costly bioinformatics solutions (that are beyond the scope of this article) to assemble very large numbers of relatively short sequence reads, and it is not yet clear that whole human genome sequences can be assembled entirely from the read and mapping information obtained by current SBS methods, though resequencing (with benefit of a reference sequence) has fallen below that price-point.

The degree to which SBS approaches can lead to further substantial cost reductions below $100,000, toward the $1,000 genome, is predicated on the degree of success of continued development and enhancement of the already commercialized systems as well as the introduction of potentially revolutionary single molecule–based SBS approaches[24] based on development by groups such as Pacific Biosciences (Menlo Park, CA, USA)[18,19,50], Helicos Biosciences (Cambridge, MA, USA)[24] and VisiGen Biotechnologies (Houston)[22]. The decisive, long-term cost bottleneck may ultimately be in the sample preparation (which requires a high degree of automation), the read-length per fragment and the cost of instrumentation (purchase price, operational lifetime and support costs)—as modulated by the instrument throughput and accuracy. It may also reside in subsequent data analysis and storage as well. Successful elimination of one or more of these bottlenecks may require the full engagement and intellectual wherewithal of the broader scientific and engineering community as its constituents recognize their collaborative role in applying their expertise to developing technologies that enable rapid, cost-effective access to DNA sequence information for a myriad of research and personalized medical uses.

ACKNOWLEDGMENT

This work was supported in part by the National Human Genome Research Institute, National Institutes of Health.

AUTHOR CONTRIBUTIONS

C.W.F. and L.R.M. wrote this review, with additions and editorial assistance from S.A.B., G.M.C., T.H., X.H., S.B.J., J.R.N., D.C.S. and D.V.V., who contributed portions of the text and read drafts of the manuscript for accuracy. J.A.S. is the scientific manager of the NHGRI Sequencing Technology Development Program; he proposed the idea for the review, provided a forum to begin its formulation at a program meeting and read the manuscript for accuracy.

COMPETING INTERESTS STATEMENT

The authors declare competing financial interests: details accompany the full-text HTML version of the paper at http://www.nature.com/naturebiotechnology/.

Published online at http://www.nature.com/naturebiotechnology/.
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/.

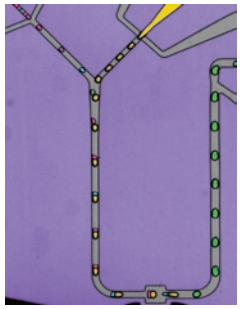
1. Schloss, J. How to get genomes at one ten-thousandth the cost. *Nat. Biotechnol.* **26**, 1113–1115 (2008).
2. Dohm, J.C., Lottaz, C., Borodina, T. & Himmelbauer, H. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res.* **36**, e105, 2008.
3. Craig, D.W. *et al.* Identification of genetic variants using bar-coded multiplexed sequencing. *Nat. Methods* **5**, 887–893 (2008).
4. Li, H., Ruan, J. & Durbin, R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.* **18**, 1851–1858 (2008).
5. Morozova, O. & Marra, M.A. Applications of next-generation sequencing technologies in functional genomics. *Genomics* **92**, 255–264 (2008).
6. Mardis, E. Next-generation DNA sequencing methods. *Annu. Rev. Genomics Hum. Genet.* **9**, 387–402 (2008).
7. Mardis, E.R. The impact of next-generation sequencing technology on genetics. *Trends Genet.* **24**, 133–141 (2008).
8. Chi, K.R. The year of sequencing. *Nat. Methods* **5**, 11–14 (2008).
9. Marguerat, S., Wilhelm, B.T. & Bähler, J. Next-generation sequencing: applications beyond genomes. *Biochem. Soc. Trans.* **36**, 1091–1096 (2008).
10. Smith, D.R. *et al.* Rapid whole-genome mutational profiling using next-generation sequencing technologies. *Genome Res.* **18**, 1638–1642 (2008).
11. Quinn, N.L. *et al.* Assessing the feasibility of GS FLX Pyrosequencing for sequencing the Atlantic salmon genome. *BMC Genomics* **9**, 404 (2008).
12. Sarin, S. *et al. Caenorhabditis elegans* mutant allele identification by whole-genome sequencing. *Nat. Methods* **5**, 865–867 (2008).
13. Holt, R.A. & Jones, S.J.M. The new paradigm of flow cell sequencing. *Genome Res.* **18**, 839–846 (2008).
14. Rothberg, J.M. & Leamon, J.H. The development and impact of 454 sequencing. *Nat. Biotechnol.* **26**, 1117–1124 (2008).
15. Kahvejian, A., Quackenbush, J. & Thompson, J.F. What would you do if you could sequence everything? *Nat. Biotechnol.* **26**, 1125–1133 (2008).
16. Shendure, J. & Hanlee, J. Next-generation DNA sequencing. *Nat. Biotechnol.* **26**, 1135–1145 (2008).
17. Branton, D. *et al.* Nanopore sequencing. *Nat. Biotechnol.* **26**, 1146–1153 (2008).
18. Levene, M.J. *et al.* Zero-mode waveguides for single-molecule analysis at high concentrations. *Science* **299**, 682–686 (2003).
19. Korlach, J. *et al.* Selective aluminum passivation for targeted immobilization of single DNA polymerase molecules in zero-mode waveguide nanostructures. *Proc. Natl. Acad. Sci. USA* **105**, 1176–1181 (2008).
20. Williams, J. System and method for nucleic acid sequencing by polymerase synthesis. US patent application US20030194740 (2003).
21. Willams, J. & Anderson, J. Field-switch sequencing. US patent application US20050266456 (2005).
22. Hardin, S., Gao, X., Briggs, J., Willson, R. & Tu, S.C. Real-time sequence determination. US patent application US20030064366 (2003).
23. Margulies, M. *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, 376–380 (2005).
24. Harris, T.D. *et al.* Single-molecule DNA sequencing of a viral genome. *Science* **320**, 106–109 (2008).
25. Milton, J. *et al.* Modified nucleotides (for polynucleotide sequencing). World and US patent application WO2004/018497, US2007/0166705 (2004).
26. Bentley, D.R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53–59 (2008).
27. 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).
28. 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).
29. 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).
30. Seo, T.S. *et al.* Four-color DNA sequencing by synthesis on a chip using photocleavable fluorescent nucleotides. *Proc. Natl. Acad. Sci. USA* **102**, 5926–5931 (2005).
31. Wu, W. *et al.* Termination of DNA synthesis by *N*6-alkylated, not 3′-*O*-alkylated, photocleavable 2′-deoxyadenosine triphosphates. *Nucleic Acids Res.* **35**, 6339–6349 (2007).
32. McKernan, K., Blanchard, A., Kotler, L. & Costa, G. Reagents, methods and libraries for bead-based sequencing. PCT patent application WO2006084132 (2007).
33. Shendure, J. *et al.* Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* **309**, 1728–1732 (2005).
34. Fuller, C.W. & Nelson, J.R. Method for nucleic acid analysis. PCT patent application WO2005123957 (2005).
35. Fuller, C.W. Rapid parallel nucleic acid analysis. US patent application US20060051807 (2006).
36. Ronaghi, M., Uhlén, M. & Nyrén, P. A sequencing method based on real-time pyrophosphate. *Science* **281**, 363–365 (1998).
37. Scheibye-Alsing, K. *et al.* Sequence assembly. *Comput. Biol. Chem.* **33**, 121–136 (2009).
38. Trapnell, C. & Salzberg, S.L. How to map billions of short reads onto genomes. *Nat. Biotechnol.* **27**, 455–457 (2009).
39. Chetverina, H.V. & Chetverin, A.B. Cloning of RNA molecules *in vitro*. *Nucleic Acids Res.* **21**, 2349–2353 (1993).
40. Church, G.M. Replica amplification of nucleic acid arrays. US patent 6,432,360 (2002).
41. 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).
42. Embleton, M.J., Gorochov, G., Jones, P.T. & Winter, G.P. *In situ* recombinant PCR within single cells US Patent 5,830,663 (1998).
43. Griffiths, A. & Tawfik, D. *In vitro* sorting method. US patent 6,489,103 (2002).
44. Holliger, P. & Ghadessy, F. Emulsion compositions. US patent 7,429,467 (2008).
45. Brenner, S. *et al. In vitro* cloning of complex mixtures of DNA on microbeads: physical separation of differentially expressed cDNAs. *Proc. Natl. Acad. Sci. USA* **97**, 1665–1670 (2000).
46. Adessi, C., Kawashima, E., Mayer, P., Mermod, J.J. & Turcatti, G. Methods of nucleic acid amplification and sequencing. PCT patent application WO2000018957 (2000).
47. Boles, T.C., Kron, S.J. & Adams, C.P. Nucleic acid-containing polymerizable complex. US patent 5,932,711 (1999).
48. Zhang, K. *et al.* Long-range polony haplotyping of individual human chromosome molecules. *Nat. Genet.* **38**, 382–387 (2006).
49. Zhang, K. *et al.* Sequencing genomes from single cells via polymerase clones. *Nat. Biotechnol.* **24**, 680–686 (2006).
50. Eid, J. *et al.* Real-time DNA sequencing from single polymerase molecules. *Science* **323**, 133–138 (2009).
51. Li, M., Diehl, F., Dressman, D., Vogelstein, B. & Kinzler, K.W. BEAMing up for detection and quantification of rare sequence variants. *Nat. Methods* **3**, 95–97 (2006).
52. Pihlak, A. *et al.* Rapid genome sequencing with short universal tiling probes. *Nat. Biotechnol.* **26**, 676–684 (2008).
53. Liu, D., Daubendiek, S.L., Zillman, M.A., Ryan, K. & Kool, E.T. Rolling circle DNA synthesis: small circular oligonucleotides as efficient templates for DNA polymerases. *J. Am. Chem. Soc.* **118**, 1587–1594 (1996).

54. Baner, J., Nilsson, M., Mendel-Hartvig, M. & Landegren, U. Signal amplification of padlock probes by rolling circle replication. *Nucleic Acids Res.* **26**, 5073–5078 (1998).
55. Fire, A. & Xu, S.Q. Rolling replication of short DNA circles. *Proc. Natl. Acad. Sci. USA* **92**, 4641–4645 (1995).
56. Blanco, L. *et al.* Highly efficient DNA synthesis by the phage phi 29 DNA polymerase. Symmetrical mode of DNA replication. *J. Biol. Chem.* **264**, 8935–8940 (1989).
57. Dean, F.B., Nelson, J.R., Giesler, T.L. & Lasken, R.S. Rapid amplification of plasmid and phage DNA using Phi 29 DNA polymerase and multiply-primed rolling circle amplification. *Genome Res.* **11**, 1095–1099 (2001).
58. Lizardi, P.M. *et al.* Mutation detection and single-molecule counting using isothermal rolling-circle amplification. *Nat. Genet.* **19**, 225–232 (1998).
59. Ramanathan, A. *et al.* An integrative approach for the optical sequencing of single DNA molecules. *Anal. Biochem.* **330**, 227–241 (2004).
60. Parameswaran, P. *et al.* A pyrosequencing-tailored nucleotide barcode design unveils opportunities for large-scale sample multiplexing. *Nucleic Acids Res.* **35**, e130 (2007).
61. Albert, T.J. *et al.* Direct selection of human genomic loci by microarray hybridization. *Nat. Methods* **4**, 903–905 (2007).
62. Porreca, G.J. *et al.* Multiplex amplification of large sets of human exons. *Nat. Methods* **4**, 931–936 (2007).
63. Kim, J.B. *et al.* Polony multiplex analysis of gene expression (PMAGE) in mouse hypertrophic cardiomyopathy. *Science* **316**, 1481–1484 (2007).
64. Williams, J.G. *et al.* An artificial processivity clamp made with streptavidin facilitates oriented attachment of polymerase-DNA complexes to surfaces. *Nucl. Acids Res.* **36**, e121 (2008).
65. Barbee, K.D. & Huang, X. Magnetic assembly of high-density DNA arrays for genomic analyses. *Anal. Chem.* **80**, 2149–2154 (2008).
66. Brenner, S. Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nat. Biotechnol.* **18**, 630–634 (2000).
67. Mujumdar, R.B., Ernst, L.A., Mujumdar, S.R. & Waggoner, A.S. Cyanine dye labeling reagents containing isothiocyanate groups. *Cytometry* **10**, 11–19 (1989).
68. Sood, A. *et al.* Terminal phosphate-labeled nucleotides with improved substrate properties for homogeneous nucleic acid assays. *J. Am. Chem. Soc.* **127**, 2394–2395 (2005).
69. Prober, J.M. *et al.* A system for rapid DNA sequencing with fluorescent chain-terminating dideoxynucleotides. *Science* **238**, 336–341 (1987).
70. Langer, P.R., Waldrop, A.A. & Ward, D.C. Enzymatic synthesis of biotin-labeled polynucleotides: novel nucleic acid affinity probes. *Proc. Natl. Acad. Sci. USA* **78**, 6633–6637 (1981).
71. Wu, W. *et al.* Termination of DNA synthesis by N6-alkylated, not 3′-*O*-alkylated, photocleavable 2′-deoxyadenosine triphosphates. *Nucleic Acids Res.* **35**, 6339–6349 (2007).
72. Canard, B. & Sarfati, R.S. DNA polymerase fluorescent substrates with reversible 3′-tags. *Gene* **148**, 1–6 (1994).
73. Tabor, S. & Richardson, C.C. A single residue in DNA polymerases of the *Escherichia coli* DNA polymerase I family is critical for distinguishing between deoxy- and dideoxyribonucleotides. *Proc. Natl. Acad. Sci. USA* **92**, 6339–6343 (1995).
74. Doublié, S., Tabor, S., Long, A.M., Richardson, C.C. & Ellenberger, T. Crystal structure of a bacteriophage T7 DNA replication complex at 2.2 Å resolution. *Nature* **391**, 251–258 (1998).
75. 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).
76. Mulder, B.A. *et al.* Nucleotide modification at the gamma-phosphate leads to the improved fidelity of HIV-1 reverse transcriptase. *Nucleic Acids Res.* **33**, 4865–4873 (2005).
77. Williams, J.G.K., Anderson, J.P., Urlacher, T.M. & Steffens, D.L. Mutant polymerases for sequencing and genotyping. US patent application US20070048748 (2007).
78. Williams, J.G.K. Polymerases with charge-switch activity and methods of generating such polymers. US published patent application US20040259082 (2004).
79. Korlach, J. *et al.* Long, processive enzymatic DNA synthesis using 100% dye-labeled terminal phosphate-linked nucleotides. *Nucleosides Nucleotides Nucleic Acids* **27**, 1072–1083 (2008).
80. Reynolds, B., Miller, R., Williams, J.G. & Anderson, J.P. Synthesis and stability of novel terminal phosphate-labeled nucleotides. *Nucleosides Nucleotides Nucleic Acids* **27**, 18–30 (2008).
81. Steffens, D.L. & Williams, J.G. Efficient site-directed saturation mutagenesis using degenerate oligonucleotides. *J. Biomol. Tech.* **18**, 147–149 (2007).
82. Ryu, J., Hong, S.S., Horn, B.K.P., Freeman, D.M. & Mermelstein, M.S. Multibeam interferometric illumination as the primary source of resolution in optical microscopy. *Appl. Phys. Lett.* **88**, 171112 (2006).
83. Mico, V. *et al.* Transverse resolution improvement using rotating-grating time-multiplexing approach. *J. Opt. Soc. Am. A Opt. Image Sci. Vis.* **25**, 1115–1129 (2008).
84. Pierce, J.R. *An Introduction to Information Theory*, edn. 2 (Dover Press, Mineola, NY, 1980).
85. Lundstrom, M. Moore's law forever? *Science* **299**, 210–211 (2001).
86. Stelzer, E.H.K. Beyond the diffraction limit? *Nature* **417**, 806–807 (2002).
87. Lloyd, S. Ultimate physical limits to computation. *Nature* **406**, 1047–1054 (2000).
88. Lloyd, S., Giovannetti, V. & Maccone, L. Physical limits to communication. *Phys. Rev. Lett.* **93**, 100501–100504 (2004).
89. Stelzer, E.H.K. & Grill, S. The uncertainty principle applied to estimate focal spot dimensions. *Opt. Commun.* **173**, 51–56 (2000).
90. Moerner, W.E. & Fromm, D.P. Methods of single-molecule fluorescence spectroscopy and microscopy. *Rev. Sci. Inst.* **74**, 3597–3619 (2003).
91. Basché, Th., Ambrose, W.P. & Moerner, W.E. Optical spectra and kinetics of single impurity molecules in a polymer: spectral diffusion and persistent spectral hole burning. *J. Opt. Soc. Am. B* **9**, 829–836 (1992).
92. Stevenson, C.L. & Winefordner, J.D. Estimating detection limits in ultratrace analysis. Part I: The variability of estimated detection limits. *Applied Spect.* **45**, 1217–1224 (1991).
93. Stevenson, C.L. & Winefordner, J.D. Estimating detection limits in ultratrace analysis. Part II: Detecting and counting atoms and molecules. *Applied Spect.* **46**, 407–419 (1992).
94. Stevenson, C.L. & Winefordner, J.D. Estimating detection limits in ultratrace analysis. Part III: Monitoring atoms and molecules with laser-induced fluorescence. *Applied Spect.* **46**, 715–724 (1992).
95. Landauer, R. Minimal energy requirements in communication. *Science* **272**, 1914–1918 (1996).
96. Zurek, W.H. Thermodynamic cost of computation, algorithmic complexity and the information metric. *Nature* **341**, 119–124 (1989).
97. Simon, S.H., Moustakas, A.L., Stoytchev, M. & Safar, H. Communication in a disordered world. *Phys. Today* **54**, 38–43 (2001).
98. Méray, L. & Demény, O. Detection limit and decision thresholds in spectrometry. *Appl. Spect.* **55**, 1102–1108 (2001).
99. Lachmann, M., Newman, M.E.J. & Moore, C. The physical limits of communication. *Am. J. Phys.* **72**, 1290–1293 (2004).
100. Gordon, J.M. & Feuermann, D. Optical performance at the thermodynamic limit with tailored imaging designs. *Appl. Opt.* **44**, 2327–2331 (2005).
101. Sheehan, P.E. & Whitman, L.J. Detection limits for nanoscale biosensors. *Nano Lett.* **5**, 803–807 (2005).
102. Prummer, M. Sick., B. Renn, A. & Wild, U.P. Multiparameter microscopy and spectroscopy for single-molecule analytics. *Anal. Chem.* **76**, 1633–1640 (2004).
103. Hubaux, A. & Vos, G. Decision and detection limits for linear calibration curves. *Anal. Chem.* **42**, 849–855 (1970).
104. Lundquist, P.M. *et al.* Parallel confocal detection of single molecules in real time. *Opt. Lett.* **33**, 1026–1028 (2008).
105. Bashford, G. *et al.* Automated bead-trapping apparatus and control system for single-molecule DNA sequencing. *Opt. Express* **16**, 3445–3455 (2008).
106. Gibson, J.D. *Principles of Digital and Analog Communications*, edn. 2 (Macmillan Publishing, New York, 1993).
107. Honigs, D. The sayings of Tomas Hirschfeld. *Applied Spect.* **40**, 11A (1986).
108. Hirschfeld, T. Limits of analysis. *Anal. Chem.* **48**, 16A–31A (1976).
109. Mouse Genome Sequencing Consortium. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520–562 (2002).
110. Eckmann, J.-P. Trading codes for errors. *Proc. Natl. Acad. Sci. USA* **105**, 8165–8166 (2008).

Imaging nascent iPS cells

Methods for generating induced pluripotent stem (iPS) cells are still inefficient, leading to heterogeneous cultures in which very few cells are fully reprogrammed. How are these rare cells to be identified? For mouse cells, this is readily achieved using genetically encoded markers and functional assays of pluripotency. For human cells, the acid test is the ability to form teratomas *in vivo*, but markers for prospective, *in vitro* identification of iPS cells have not been validated. Schlaeger, Daley and colleagues use automated fluorescence microscopy to follow the expression of several markers during the reprogramming of human fibroblasts over ~3 weeks. Analysis of colonies that are morphologically similar to embryonic stem cell colonies reveals three distinct colony types. Only one type of colony—which is $CD13^{-}$, GFP^{dim}, $SSEA\text{-}4^{+}$, $TRA\text{-}1\text{-}60^{+}$, $Hoechst^{dim}$ and can be expanded into a cell line—proves to be fully reprogrammed. In contrast, the other two colony types seem to be largely trapped in partially reprogrammed states. This study defines a marker signature that allows prospective identification of bona fide human iPS cells. It also invalidates certain markers previously used in isolation to score iPS cells, such as SSEA-4 and alkaline phosphatase. [**Letters, p. 1033; News and Views, p. 997**] *KA*

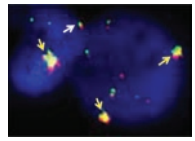
Microdroplet genome enrichment



Enriching and sequencing specific regions of the genome is a promising approach for studying genetic variation in a population. Frazer and colleagues demonstrate that using microdroplet PCR to enrich targets of interest may offer advantages over approaches involving traditional PCR, oligo capture and molecular inversion 'padlock' probes. In microdroplet PCR, microfluidics enables thousands of genomic regions to be amplified in parallel in tiny droplets, each containing a distinct pair of PCR primers. In evaluating enrichment technologies, key performance parameters include the evenness of coverage and targeting specificity; both are indicators of efficient use of sequencing resources, which enables more samples to be processed and greater statistical power to be achieved in population-based studies. Frazer and colleagues demonstrate multiplexed enrichment using ~4,000 primer pairs to obtain 1.35 Mb of sequence from 7.5 µg of starting DNA with competitive specificity and coverage performance. The use of PCR also enables enrichment of genes with high nucleotide similarity, such as those within a gene family. Further optimization to allow nanogram quantities of DNA to be enriched using ~20,000 primer pairs may soon make this the method of choice for targeted genome sequencing. [**Research Articles, p. 1025; News and Views, p. 998**] *CM*

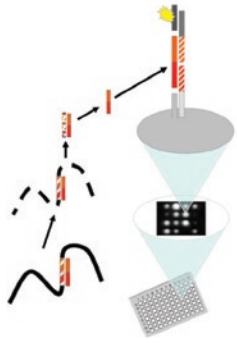
Written by Kathy Aschheim, Laura DeFrancesco, Markus Elsner, Michael Francisco, Craig Mak & Lisa Melton

Prioritizing cancer fusions



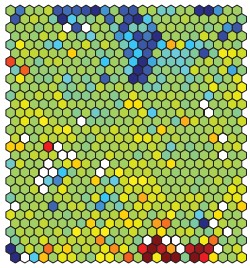
Aberrant gene fusions are often found in the genomes of cancer cells. Most fusions are by-products of tumor progression, but a few, it is thought, actually drive some cancers. Chinnaiyan and colleagues demonstrate methods for prioritizing likely 'driver' mutations, a challenging task that is important in light of the many fusions that are typically identified by high-throughput array- and sequencing-based technologies. The method of Wang *et al.* combines insights into patterns of chromosomal rearrangements that generate functionally important fusions and bioinformatic analyses of the fusion partners. The researchers analyze a database of known cancer fusions and observe that a given gene may be involved in many fusions but with different partners that tend to interact physically with a common third gene and participate in specific cellular processes. Chinnaiyan and colleagues apply this knowledge to prioritize fusions in paired-end next-generation sequencing data of 12 lung adenocarcinoma cell lines. The top-scoring hit in one cell line was a fusion involving the transcription factor NFE2, which is shown to influence cancer cell proliferation and invasion, and found to be a recurrent rearrangement in patient samples. [**Analysis, p. 1005**] *CM*

Expression profiling of fixed cells



Isolating subsets of cells in heterogeneous tissues cannot always be done with antibodies against cell-surface markers and may require intracellular staining. In these cases, it is not possible to study gene expression in the sorted cells by conventional methods like RT-PCR or northern blot analysis owing to the cross-linking caused by fixation. Working with subsets of mouse pancreatic islet cells, Pechhold and colleagues show that the quantitative nuclease protection assay allows transcriptional analysis of fixed cells purified by fluorescence-activated cell sorting (FACS). The assay is thought to perform well in the presence of cross-linking because it relies on mRNA capture by short 50-mer cDNA probes. The authors apply the method to gain new insight into islet-cell subsets, including the finding that beta cells in pregnant mice express Mafb, a gene believed to be expressed only in the developing pancreas. The ability to monitor gene expression in fixed, FACS-purified cells should facilitate the study of heterogeneous tissues, including diseased tissues such as tumors. [Letters, p. 1038] *KA*

Drug metabolizers screened



Many drugs are metabolized in the liver, where they are substrates of the cytochrome P450 enzymes. To better understand the chemical preferences of different cytochrome P450 isozymes, Auld and colleagues determine potency values for five of the most important isozymes against a diverse chemical library of >17,000 compounds. Their screening library includes ~1,100 FDA-approved drugs and thousands of small molecules representative of drug discovery libraries. The screening results, which represent the largest publicly available resource of its kind, reveal a number of interesting facts about the isozymes and the compounds tested. The FDA-approved drugs tend to show low activity against specific isozymes, suggesting that studies of these individual isozymes may help optimize new drugs. To shed some light on adverse drug metabolism, Auld and colleagues also identify chemical structures that confer selectivity towards specific isozymes. This public database should enable future development of predictive models of cytochrome P450 activity and may guide the use of *in vitro* P450 assays in early-phase drug development. [**Resource, p. 1050**] *CM*

Organization of the *E. coli* genome

Even a decade after the first complete sequence of an *Escherichia coli* genome was published, we still lack a complete description of its functional organization. In particular, we lack a full understanding of which elements of the genome can be transcribed into mRNA and translated into proteins and how different sequences are combined under different circumstances to form RNA molecules. Palsson and colleagues use high-throughput, genome-wide measurements of RNA polymerase binding regions, mRNA abundance, 5' sequences and translation into proteins to provide a more complete annotation of the *E. coli* genome. They define 3,138 co-transcribed sequences—so-called modular units that comprise parts of the genome that are always transcribed into mRNA together with one or more transcriptional start sites. They further analyze which start sites are chosen and how different modular units are combined to form a single mRNA under different growth conditions. The authors identified 4,661 different combinations, called transcriptional units, consisting of one or more modular units and a single transcriptional start site in the four growth conditions analyzed. This extensive experimental annotation of the *E. coli* genome will enable better models of gene regulation to be developed. [**Resource, p. 1043**] *ME*

Throwing down the gauntlet

Since the completion of the human genome, attention has increasingly focused on improving sequencing technology based on approaches using DNA polymerase or DNA ligase. Already these efforts have spawned a vibrant sector, and at least seven instruments are commercially available that perform next-generation sequencing on a large scale using sequencing by synthesis. Yet, the cost has not broken the $1,000 barrier—a goal of the National Human Genome Research Institute of the US National Institutes of Health—or even the $100,000 barrier. Fuller and colleagues explain some of the reasons why, and detail the challenges facing developers of sequencing by synthesis technology, step by step, from sample preparation to modes of detection. The authors believe that a cross-disciplinary effort that encompasses biochemistry, chemistry, physics and engineering is needed to take on these challenges. [**Review, p. 1013**] *LD*

Patent roundup

In the debate about accessing and sharing the benefits of biological resources, patents may not be establishing clear property and use rights, and this may promote biodiversity destruction and decline. According to Lawson, the challenge in implementing any new patent arrangements will be in adjusting the schemes for patents and other forms of intellectual property to suit conservation and sustainable use objectives. [**Patent Article, p. 994**] *MF*

Recent patent applications in biological imaging. [**New Patents, p. 995**] *MF*

A federal court has ruled in favor of diagnostic test maker Prometheus Labs (San Diego), ending nearly a year of uncertainty over the patentability of medical diagnostic and treatment methods. [**News in Brief, p. 963**] *LM*

Next month in nature biotechnology

- A biofuel generated by photosynthesis
- Measuring alternative lengthening of telomeres
- Regulatory elements at nucleotide resolution
- Polypeptide-mediated extension of protein half-life