



ELSEVIER

Sequencing single molecules of DNA

Hagan Bayley

In 2004, the NIH set a remarkable challenge: the \$1000 genome. Roughly speaking, success would provide, by 2015, the ability to sequence the complete genome of an individual human, quickly and at an accessible price. An intermediate goal of a \$100 000 genome was set for 2010. While the cost of Sanger sequencing has dropped dramatically over the past two decades, it is unlikely that the \$100 000 genome will be achieved by this means. New massively parallel technologies will push the cost of sequencing towards this mark, but it is doubtful whether these efforts will match the \$1000 goal. The best bets for ultrarapid, low-cost sequencing are single-molecule approaches.

Addresses

Department of Chemistry, University of Oxford, Chemistry Research Laboratory, Mansfield Road, Oxford, OX1 3TA, UK

Corresponding author: Bayley, Hagan (hagan.bayley@chem.ox.ac.uk)

Current Opinion in Chemical Biology 2006, 10:628–637

This review comes from a themed issue on
Biopolymers
Edited by Hagan P Bayley and Floyd Romesberg

Available online 20th November 2006

1367-5931/\$ – see front matter
© 2006 Elsevier Ltd. All rights reserved.

DOI 10.1016/j.cbpa.2006.10.040

Introduction

Remember that time is money

Benjamin Franklin: *Advice to a young tradesman*, 1748

Given the various approaches to sequencing and the different outcomes that might be envisaged, the \$1000 genome cannot be precisely defined. Although worthy attempts at clarity have been made [1] (<http://tinyurl.co.uk/8qwu>), issues involving *de novo* sequencing versus re-sequencing, haplotyping and acceptable error rates are not readily settled. But, in this short review, let's not be too fussy, and simply define a \$1000 genome as the medically useful (~ 1 in 10 000 finished error rate) re-sequenced genome of an individual human being, obtained at affordable cost (i.e. well less than 1% of the total lifetime expenditure on that individual's health care) ([1]; <http://tinyurl.co.uk/8qwu>). Personal genome sequences obtained in this way, perhaps at birth, would facilitate disease prevention, improve diagnosis and guide better treatment. The availability of a large number of

human genome sequences would greatly facilitate the investigation of comparative human genomics. Cheap, fast sequencing technology would further medical research, for example, by allowing the analysis of tumor progression at an unprecedented level of detail.

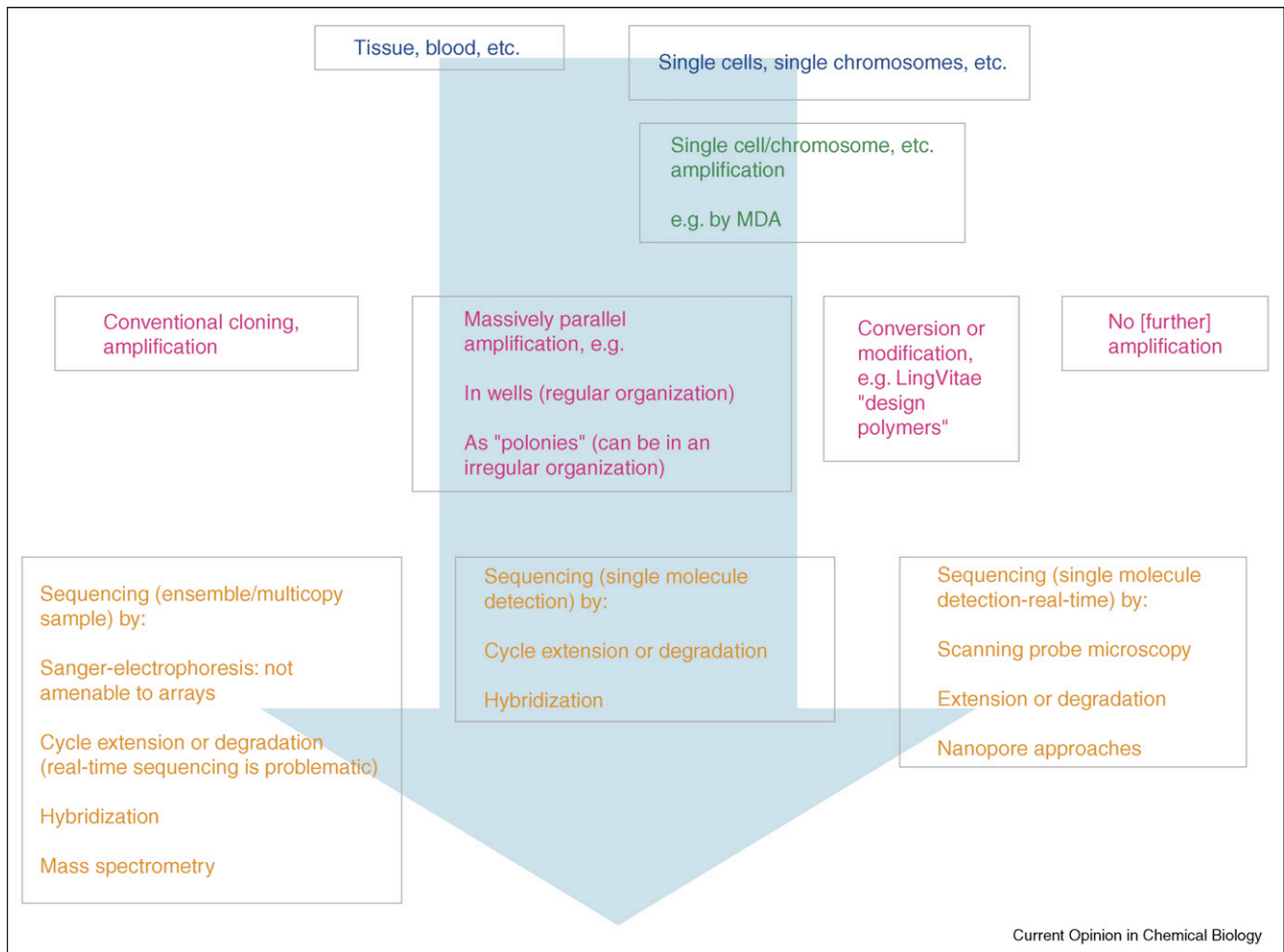
The scale of the problem is daunting. The diploid human genome contains 6000 million bases, roughly the length of 1000 Russian novels. The top of the range, Applied Biosystems 3730x/ sequencer uses 96 capillaries and produces about 18 bases of raw read per second. One hundred of these machines could then sequence a genome with about $5\times$ coverage in roughly one year, assuming negligible additional non-parallel time for sample preparation and sequence assembly. And, not unrelatedly, it would be an expensive business: at present costs, more than \$30 million.

At the inception of the human genome project in 1984, the cost of DNA sequencing by the Sanger chain termination method was about \$1 per raw base and the overall cost of the human genome project is often given as \$3 billion, although this is of necessity a rough estimate and finished sequence continues to be reported. By automation, miniaturization, improved biochemistry, better computer software and hardware for sequence assembly and local and collaborative management techniques, the costs have been driven down by ~ 1000 -fold. Several groups are attempting to push down the costs of Sanger sequencing yet further (for active NIH grants see: <http://www.genome.gov/10001799>). Notably, the Mathies group at Berkeley is implementing a laboratory-on-a-chip approach for amplifying, purifying and sequencing DNA in a single device [2*]. This remarkable effort has resulted in a huge reduction in reagent consumption, but it would be surprising if the costs could be brought below the \$100 000 mark for a complete genome sequence.

Low cost sequencing: general considerations

Below, we discuss new highly parallel approaches to DNA sequencing and then follow on with more detail about single molecule approaches. A brief consideration of important general issues concerning DNA sequencing serves to clarify the discussion (Figure 1). First, a source of genomic DNA is required. For personal genome sequencing, this is likely to be from leukocytes separated from peripheral blood, or where less DNA can be used from a buccal swab or a hair follicle. For ensemble (multi-copy) sequencing, DNA is extracted from the whole sample. In other cases, for example, to monitor individual cells in tumors or for haplotype mapping, a second stage, multiple displacement amplification (MDA) on a single

Figure 1



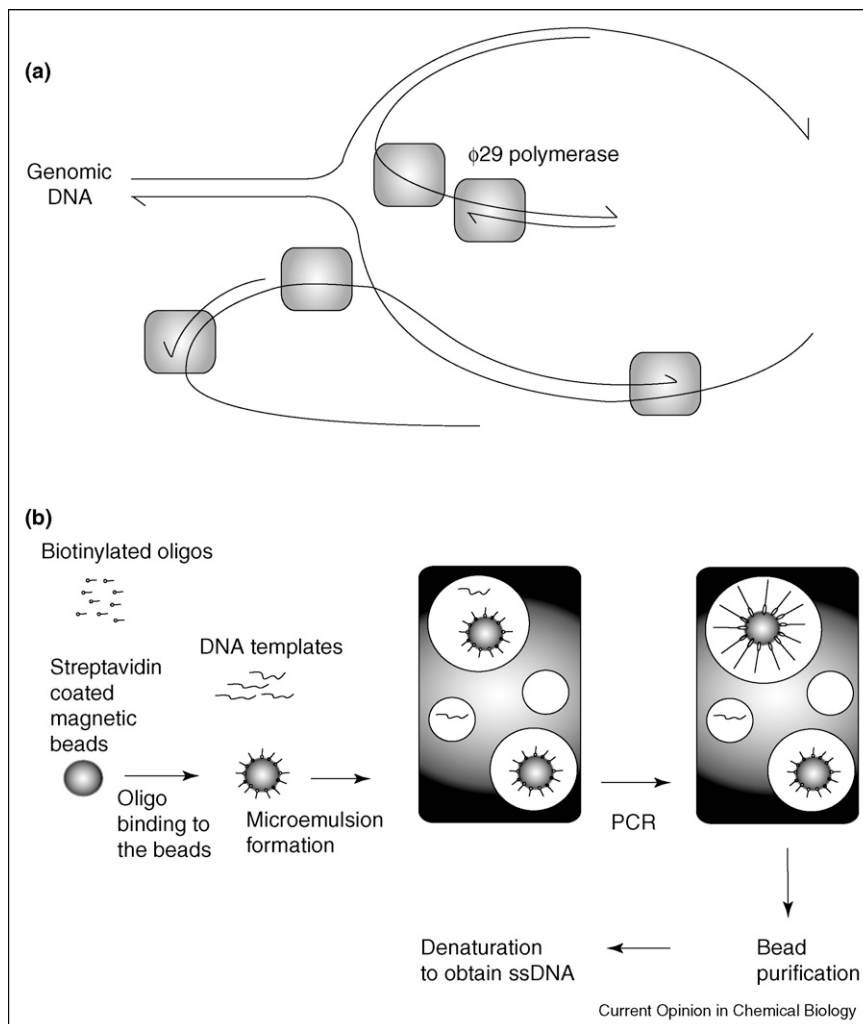
Routes to genomic sequence. Blue text, source of DNA; green text, amplification of a single DNA copy; pink text, DNA amplification and/or modification; orange text, sequencing.

cell or a single chromosome might be required in which isothermal strand-displacement amplification is performed with an enzyme such as $\phi 29$ DNA polymerase (Figure 2a). The considerable potential of this approach was demonstrated by Zhang and colleagues who evaluated variations of the procedure for sequencing entire bacterial genomes from single cells [3[•]]. This form of amplification is better able to deal with repeats than PCR, and the error rate of < 1 in 2×10^5 is well below the acceptable error rate for sequencing. It is likely that MDA will be further refined in the coming years, so that it can be routinely applied to single human cells and chromosomes.

A third stage in sequencing (Figure 1) is the amplification of discrete segments of DNA for sequence determination. For Sanger sequencing, conventional cloning or PCR amplification is a familiar necessity. Recently, massively parallel amplification techniques have been tested on bacterial genomes. In two notable cases, polymerase

colonies (colonies) were formed [4^{••},5^{••}]. In this approach, individual genomic fragments with or without prior amplification are dispersed and then amplified in such a way that the products remain separated. There are various ways to accomplish this but in both of the recent papers emulsion PCR and bead capture were used [6] (Figure 2b). The DNA fragments are amplified within water-in-oil emulsion droplets and captured within the droplets by hybridization to beads for subsequent sequencing. DNA for sequencing can also be converted or modified in various ways. In one interesting conversion, the company LingVitae has used an enzymatic process, carried out in a microfluidics system, to convert each base to a binary combination of two 10-base segments (<http://www.lingvitae.com>). These 'design polymers' might be useful substrates for single-molecule sequencing (see below). DNA with fluorescently labeled bases suitable for single-molecule exonuclease sequencing (see below) can be prepared by transcription with a polymerase, but so

Figure 2



Advances in DNA amplification. **(a)** Multiple displacement amplification (MDA) of DNA from a single cell or a single chromosome. The potential of the approach for genome sequencing has been demonstrated for individual bacterial cells [3]. Isothermal strand-displacement amplification of genomic DNA is carried out with an enzyme such as $\phi 29$ DNA polymerase. **(b)** Emulsion PCR. Individual genomic DNA fragments are amplified within water-in-oil emulsion droplets, captured within the droplet by hybridization to a bead and converted by denaturation to tethered single-stranded DNA for sequencing (e.g. Figure 3) [4^{**},5^{**},6].

far it has not been possible to obtain complete substitution, although significant technological improvements are being made [7,8]. The modification of bases might also be achieved post-transcriptionally after placing orthogonal chemical 'hooks' in the DNA by using minimally modified bases [9] (<http://www.genome.gov/10001799>).

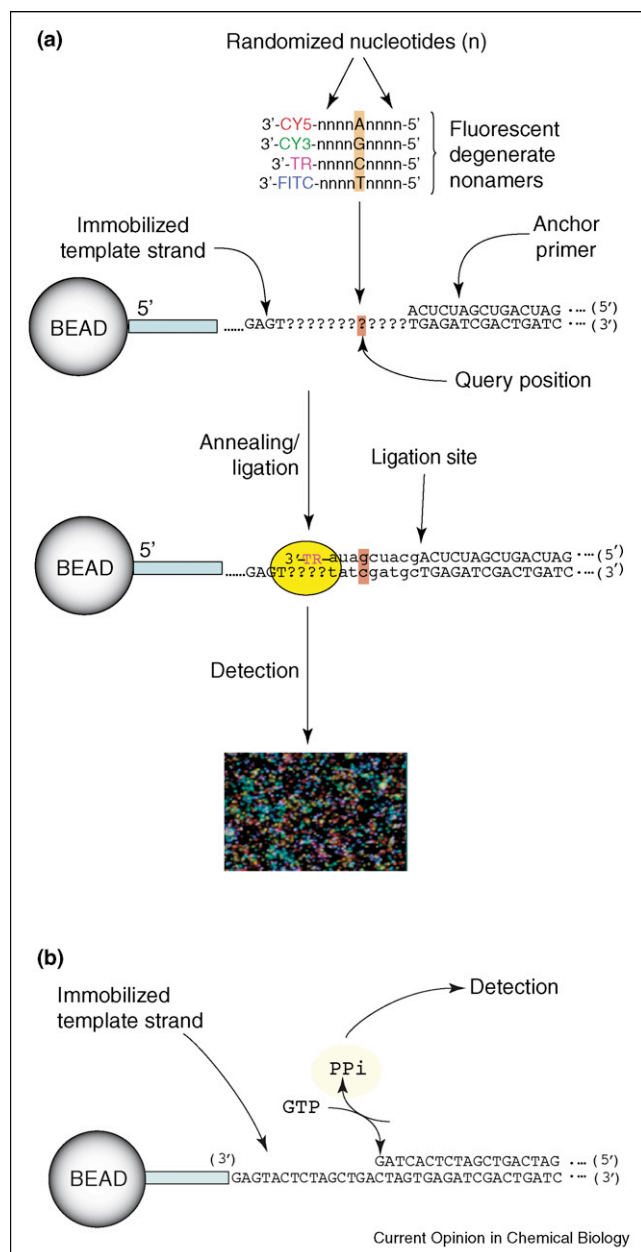
Fourth and finally, DNA sequencing is conducted, either by ensemble-level detection, by using single-molecule detection implemented in a cyclic fashion, or by single-molecule detection implemented in real time (Figure 1).

Recent advances in ensemble sequencing

Two recent papers illustrate the power of massively parallel amplification, followed by non-electrophoretic

sequence determination [4^{**},5^{**}] (Figure 3). Both feature emulsion PCR (Figure 2b). In the first case, polony beads were immobilized in polyacrylamide in an irregular but immutable fashion [4^{**}]. A limited stretch of sequence was then determined in parallel for each of over one million beads by automated cycles of a sequential hybridization and ligation designed to test in turn 26 individual positions and report the identity of the base with a corresponding fluorophore (Figure 3a). Sequence determination was monitored with an inexpensive fluorescence microscope. Two 60 h instrument runs were required for an entire *Escherichia coli* genome, providing 71% high quality coverage (4 \times). The cost of \$0.00011 per base was roughly one-ninth that of conventional sequencing.

Figure 3



Recent advances in massively parallel DNA sequencing. In both cases, fragments of bacterial genomic DNA were amplified by emulsion PCR (Figure 2b). **(a)** Shendure and colleagues sequenced over one million DNA strands per run by the hybridization/ligation of fluorescent degenerate nonamers [4**]. A field of hybridized beads from one of the 26 sequencing cycles is shown. **(b)** Margulies and colleagues sequenced an average of 110 bases for hundreds of thousands of DNA strands per run by using a pyrosequencing protocol with each bead in a picoliter well [5**].

In the second approach, developed by 454 Life Sciences, the beads were placed in individual 75 μ l wells in a fiber-optic slide and examined by a cyclic pyrosequencing protocol in which the liberated pyrophosphate is detected by a coupled fluorescence assay when the correct base is incorporated by a polymerase at a specific position [5**].

(Figure 3b). Over 100 bases of sequence could be determined for each bead. 25 million bases, at 99% or better accuracy, were revealed in a single 4 h run. In one such run, the small *Mycoplasma genitalium* genome (0.5 Mb) was determined with 96% coverage at 99.96% accuracy. The costs of this approach were not described in detail, but they are likely to be similar to the hybridization and ligation procedure. If the costs were pushed down another 10-fold, these technologies would approach the \$100 000 mark. At the same time, various problems would have to be ironed out, such as the difficulty of dealing with repeat sequences and homopolymer runs. The obstacles are under active investigation, for example the use of reversible terminators for pyrosequencing would help cope with homopolymers [10,11].

We can expect further variants of highly parallel sequencing. For example, Solexa have devised an ingenious technique for on-chip amplification of individual dispersed DNA fragments and combined it with a polymerase-based 'sequencing by synthesis' procedure that uses reversible chain termination with fluorescent nucleotides (www.solexa.com). All four bases are presented in each cycle, reducing incorporation errors, and when the 3'-hydroxyl is released for the subsequent cycle, the fluorophore is also removed. Alternative approaches for monitoring base incorporation during sequencing by synthesis are under development. For example, incorporation can be detected when templates are immobilized on the surface of a field-effect transistor, based on the additional negative charge of the extended strand [12*]. So far, this approach has been limited to < 10-base reads and the expected difficulties with homopolymer runs are yet to be tackled effectively.

Here, we have omitted discussion of alternative approaches such as sequencing by hybridization (the hybridization of fluorescently labeled genomic fragments to oligonucleotide arrays) [13] and mass spectrometry [14] (Figure 1), which might be best suited to the rapid resequencing of specific genomic sub-sequences, rather than complete genome sequencing or resequencing.

Single-molecule sequencing

By single-molecule sequencing we mean any technique that is capable of sequencing one DNA molecule at a time, whether that molecule is a constituent of an amplified sample or unamplified DNA (Figure 1). Single-molecule sequencing has technical advantages that remain useful for DNA obtained by cloning or PCR. It might also be applied to unamplified DNA from a relatively plentiful (multicopy source), amplified DNA from a single cell (e.g. obtained by MDA) and unamplified DNA from a single cell, although a complete genome sequence from the latter would be extremely difficult to obtain.

The primary advantages of single-molecule sequencing are twofold. First, individual fragments of DNA need not

be amplified, or when they are far fewer cycles of amplification are required. Second, sequencing can be in real-time (i.e. there need not be protracted cycles of hybridization or successive enzymatic steps). This is because, in the case of single-molecules, real-time sequencing procedures cannot become dephased, e.g. exonuclease sequencing (see below) would be impracticable on an ensemble.

Although the entire area of single-molecule sequencing remains speculative, fluorescence-based enzymatic sequencing and nanopore sequencing have emerged as leading contenders in the race to the \$1000 genome. Scanning probe techniques are also under investigation. A likely ultimate speed for nanopore sequencing is 1 ms per base, whereas real-time enzyme-based sequencing is likely to be about ten times slower on the basis of typical turnover numbers. A single nanopore would then sequence a genome's worth of DNA in about 70 days, and a 1000 pores working in parallel would give $10\times$ coverage in less than a day. Similarly, Church and colleagues have estimated that a machine for genome re-sequencing with the same unit cost as a typical multi-channel capillary electrophoresis sequencer would need to perform 60-base reads with 99.7% accuracy at $500\,000\text{ s}^{-1}$ [1]. A 2 Mb bacterial genome could be sequenced in 20 s by such an instrument!

Recent advances in single-molecule sequencing

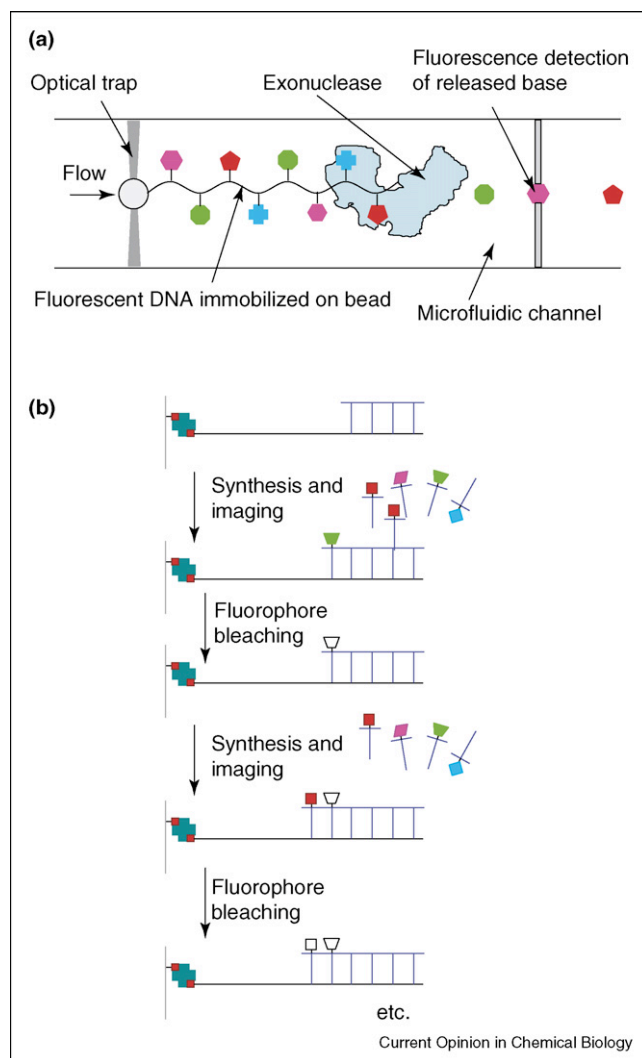
Scanning probe techniques

The proposal that DNA might be sequenced by electron microscopy was made decades ago. Nowadays, various scanning probe microscopies are candidate technologies, but no useful results have been forthcoming. In the related area of force microscopy, Lindsay and colleagues at Arizona State University are attempting to use AFM tips to pull DNA through surface-anchored rings such as cyclodextrins (<http://www.genome.gov/10001799>). The idea is that each base would offer a characteristic resistance as it passes through the ring. It is too early to evaluate this technology.

Fluorescence techniques: exonuclease sequencing

Exonuclease sequencing was conceived by Keller and colleagues in the late 1980s. The idea is to transcribe DNA with a polymerase that allows the introduction of four fluorescent nucleotides, with one corresponding to each type of base in the template strand. A single molecule of the product DNA is then anchored in a device, such as a capillary with buffer flow, and digested with an exonuclease. As each base is released, it is identified downstream in real time by single-molecule fluorescence (Figure 4a). At the single-molecule level, sequencing will remain in phase. With ensembles, dephasing is inevitable (e.g. a nuclease may be at position 17 on one strand, while another has reached position 19). Exonucleases do not have interfering

Figure 4



Approaches aimed at single-molecule DNA sequencing. (a) Exonuclease sequencing. Enzymatic digestion is used to cleave one base at a time from a transcript substituted with fluorescent nucleobases. Each base is identified by single-molecule fluorescence spectroscopy [15]. (b) Sequencing by synthesis at the single-molecule level. The enzyme-catalyzed addition of each base is monitored by a fluorescence signal (e.g. by using nucleotides with fluorescent bases [16]).

secondary activities and the approach is amenable to highly parallel implementation. There are two significant obstacles: achieving 100% substitution with fluorescent bases [7,8]; and fluorescent impurities [15]. Given this, a major advance would be the ability to detect individual unmodified nucleoside monophosphates.

Fluorescence techniques: sequencing by synthesis

A strong effort is being placed on sequencing by synthesis at the single-molecule level in which the enzyme-catalyzed addition of each base is monitored by a fluorescence signal. This approach can be implemented in a highly parallel fashion at the ensemble level by cycle sequencing

(see above). But implementation at the single molecule level would ultimately allow real time sequencing, again because there would be no dephasing (Figure 4b).

Quake's laboratory reported the repeated incorporation by a DNA polymerase of fluorescent nucleotides into individual DNA strands on a chip with single base resolution, which allowed the determination of sequences of up to 5 nt [16]. Background fluorescence was reduced by using a protocol involving FRET between the incorporated nucleotides. The complex procedure involved the cyclic introduction of reagents and a photobleaching protocol to eliminate fluorescence from incorporated nucleotides, so that subsequent incorporation events could be observed. Helicos is now pursuing this work to produce a practicable sequencing technology.

Pacific Biosciences is pursuing technology developed at Cornell University, in which polymerase reactions can be carried out in zeptoliter observation volumes [17]. This permits the use of high concentrations of fluorescent nucleotides and the selective observation of fluorescence associated with an immobilized enzyme. Incorporation events were observed as bursts of fluorescence followed by photobleaching. The technology is being modified so that it can be used with nucleotides tagged on the γ -phosphate from which the fluorescent label is released upon base incorporation into the growing strand (<http://www.genome.gov/15015202>). There are several related technologies under investigation and it is possible to mix-and-match the underlying concepts. A bead-based single-molecule approach using fluorescent nucleotides, modified on the γ -phosphate, is being developed by LiCor (<http://www.genome.gov/12513162>). Visigen are using FRET between an immobilized fluorescent donor-tagged polymerase and γ -labelled nucleotides (<http://www.genome.gov/15015202>).

At less than 10 bases, current read lengths for single-molecule fluorescence sequencing are not impressive and therefore a great deal of developmental work continues to be focused on engineered polymerases and tagged nucleotides, both of which would also be useful in ensemble cycle sequencing. Better polymerases would have reduced exonuclease activity, an improved ability to handle unusual nucleotides and reduced cycle times. Better nucleotides would be free of trace impurities, they might mediate reversible chain termination (i.e. contain protected 3' positions) and use dyes that are readily chemically or photochemically cleaved or bleached.

Nanopore approaches

Over a decade ago, it was proposed that DNA might be sequenced by pulling a single strand electrophoretically through a nanopore, specifically a protein pore. At the same time, an ionic current would be driven through the pore by the applied potential and, if each base that passed

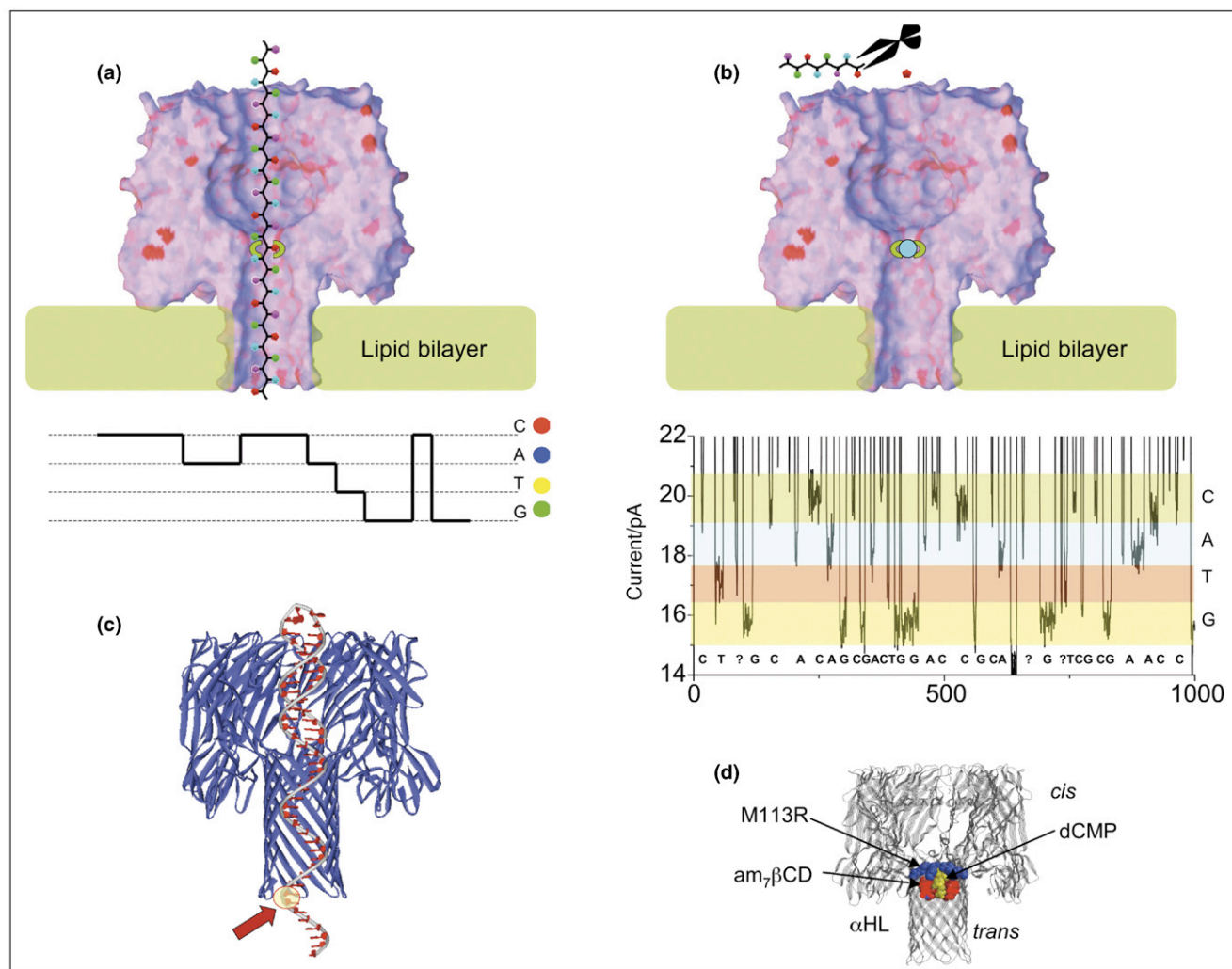
modulated the current in a way that depended upon its identity, the DNA could be sequenced (Figure 5) [18,19]. In addition to the advantages of single-molecule sequencing that have already been discussed, nanopore sequencing would be reagent free and very rapid.

Although discrimination between homopurine and homopyrimidine sequences was achieved early on with the α -hemolysin pore [20], and a wide variety of additional interesting experiments with threaded DNAs have been performed, researchers in this area have not yet come close to sequencing DNA. In all these experiments, the wild-type α HL pore was used. There are two fundamental issues that must be addressed: first, molecular recognition of the bases as they pass through the pore; and second, a reduction in the rate of DNA transport through the pore to a speed at which this recognition can be registered. These issues can be addressed through protein engineering.

Recognition might be established by two factors: the dwell time of the base at the recognition site and the effect of the base on the current amplitude. Even if the mean dwell times for the four bases could be separated by four orders of magnitude, there would be considerable overlap between individual dwell times and therefore recognition must be based primarily on current flow, perhaps with additional information from individual dwell times. Two recent papers have examined the issue of recognition. In the first, α HL•DNA pseudorotaxanes were made by threading ssDNAs with bulky terminal hairpins into the α HL pore. The hairpins lodged at the internal constriction and held the DNAs stationary within the pore (Figure 5c). By probing the system with a collection of DNAs, it was demonstrated that it is possible to determine the position and identity of an A base within a long run of Cs from the extent of current block [21*]. In this way, both important aspects of recognition, single nucleobase resolution and nucleobase discrimination, were demonstrated for a stationary nucleic acid. Remarkably, this was done with the wild-type protein. Further, the recognition site appeared to be at the *trans* mouth of the pore (Figure 5c), providing a useful clue for future studies with engineered pores. A second study, in which individual nucleoside monophosphates were analyzed by using an engineered pore containing a cyclodextrin adaptor, showed that all four bases can be recognized at the single molecule level [22*], again demonstrating nucleobase discrimination and providing the basis of a label-free detection method for exonuclease sequencing (Figure 5d).

At the transmembrane potentials required for threading and transport, DNA moves through the α HL pore rapidly, at 1 to 5 μ s per base, rather too fast for single nucleotide resolution with conventional recording procedures; there is too much noise at the required bandwidth to distinguish between G, A, T and C, even if it is assumed that the

Figure 5



Sequencing single molecules of DNA with engineered protein nanopores. Two approaches to reagent-free nanopore sequencing: **(a)** the direct sequencing of single-stranded DNA; **(b)** exonuclease sequencing with direct identification of the released bases. **(c)** By using DNAs with hairpins that become immobilized within the α HL pore, Ashkenasy and colleagues have shown that single nucleobase resolution and nucleobase discrimination can be achieved [21*]. **(d)** Astier and colleagues have shown that individual nucleoside monophosphates can be identified with an engineered α HL pore equipped with a molecular adaptor [22*]. The trace in (b) is actual single-molecule detection data from a mixture of all four dNMPs.

amplitude difference between the four states has been optimized. Therefore, the rate of DNA transport must be slowed or the recording technology further optimized. The issues involved in the latter are understood and progress has been made [23]. It is worth noting that if a time resolution of 10 μ s per base could be achieved, no enzyme that might be used for single-molecule sequencing could compete. In parallel, attempts are being made to slow down DNA transport through the α HL pore by a variety of protein engineering strategies including the use of molecular brakes (e.g. short covalently attached oligonucleotides), the provision of a high internal viscosity (e.g. with covalently attached polymers), and the attachment of nucleic acid-handling enzymes. It might also be possible to lower DNA into a pore with an AFM tip [24] or

to use a magnetic field to slow down DNA derivatized with a magnetic bead [25].

Protein pores versus non-protein pores or gaps

At least 10 research groups are examining non-protein pores in experiments aimed at DNA sequencing. It is interesting to compare the potential advantages of these structures with protein pores. First, it has been suggested that non-protein pores are more robust than protein pores. But here, it should be recognized that protein pores are stable enough for DNA sequencing. α -Hemolysin and related β -barrel proteins continue to function normally in lipid bilayers at close to 100 °C [26*]. Therefore, it will be possible to carry out protein-based nanopore sequencing

under conditions that denature DNA but not the nanopore protein. Second, non-protein nanopores do not require a lipid bilayer. This is an asset, but lipid bilayers are also stable at high temperatures. The susceptibility of bilayers to mechanical shock is being addressed by polymerization, aperture design and physical support. Recent contributions include a new investigation of polymerizable bilayers [27^{*}] and improved S-layer (porous 2D protein crystal) supports for lipid bilayers [28]. It might even be possible to forgo the bilayer with a hybrid system where a protein pore is contained within a slightly larger microfabricated aperture. It has also been suggested that it will be easier to make arrays of non-protein pores. This might be true if the failure rate for aperture formation is low (e.g. an ion beam could be scanned over the substrate). At the same time, improved approaches for introducing single copies of membrane proteins into lipid bilayers have been developed by using physical transfer with plastic or glass probes [29,30^{*}]. Finally, in non-protein pores DNA bases might be distinguished by means that do not require molecular recognition. For example, as a base passes through a pore or gap, a transverse tunneling current might be measured [31] or the local capacitance of the system determined [32]. A potential advantage of these concepts, which have not yet been reduced to practice, is that they might not require a perfect aperture. This is useful because, at the present state of the art, there is considerable difficulty in making non-protein pores or gaps reproducibly and there are few data on their stability once prepared.

General issues concerning non-protein and protein nanopores

There are several issues that non-protein and protein nanopores have in common. First, they have the potential to perform very long reads, far longer than the 700 nt or so usually obtained by Sanger sequencing. Disease-related genetic variation in human populations is often monitored through single-nucleotide polymorphisms (SNPs). These studies can be enhanced by knowing whether SNPs are present on the same chromosome (haplotyping). Because interesting SNPs can be quite widely separated, a method that can read tens of kilobases of sequence would be extremely useful [33].

The rate of nanopore sequencing would be greatly reduced if there were long pauses between the entry of DNA strands. Detailed studies of the conditions that increase the rate of DNA threading are required. For example, because the rate of threading increases exponentially with the applied potential [34], it might be useful to thread at the highest potential possible and then quickly back off by using a feed-back mechanism [35].

Nanopore sequencing might profit from the ability to convert the DNA to 'design polymers' for improved detection. In this process, which is practised by the

company LingVita, each base of a dsDNA is converted to a specific 20 bp sequence by a series of enzymatic reactions (<http://www.lingvita.com>). These sequences should be more readily recognized during nanopore sequencing than individual bases. Further, the design polymer segments can be derivatized with bulky groups to provide even greater distinction. The difficulty of placing unnatural nucleotides in DNA strands by transcription with a polymerase has already been noted. Both design polymers and post-synthetically modified polymers might also be useful for exonuclease sequencing.

However, in the long term, if nanopore sequencing is to outcompete massively parallel sequencing methods, it seems likely that direct sequencing without any amplification or base modification will be essential. Besides their expense, cloning and amplification procedures are prone to errors and bias. Ideally, DNA for nanopore sequencing would be obtained by minimal treatment of extracts from a few million cells, which would be available from, say, a buccal swab. If nanopore sequencing can be performed without amplification, it opens up possibilities such as the direct sequencing of alternatively spliced mRNAs or the analysis of methylated genomic DNA.

Alternative approaches to sequencing with nanopores

Nanopores might be used to facilitate sequencing carried out by using principles other than the direct reading of ssDNA. It has been shown that protein pores can be used in stochastic sensing [36] to detect DNA by hybridization [37]. In this case, hybridization is not driven to completion, rather individual association and dissociation events are observed. Although nanopore sequencing by this means may not be a practical approach, it is possible that other stochastic detection procedures based on hybridization could work; for example, the probing of individual immobilized genomic DNA fragments on a chip with a series of fluorescent oligonucleotides. The label-free detection with an engineered protein nanopore of individual nucleoside monophosphates was described earlier [22^{*}] (Figure 5b,d). If exonuclease sequencing is to make use of this finding, all the released bases must be fed into the pore and exit from the opposite side, so that the sequentially released bases do not interfere with one other [22^{*}]. It seems likely that this will require the attachment of an exonuclease at one of the openings of the pore.

Conclusion

New massively parallel technologies are quickly pushing the cost of genome sequencing towards the \$100 000 mark, but these approaches will soon reach limits set by reagent costs and cycle times. At the same time, efforts in single-molecule sequencing, including nanopore sequencing, are picking up pace and recent results suggest that a

\$1000 genome may be attained in this way within the next decade.

Acknowledgements

Work on DNA sequencing in the Ghadiri and Bayley laboratories is supported by the NIH. Yann Astier, Orit Braha, Rethi Madathil and Giovanni Maglia provided figures and helpful comments.

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- of special interest
 - of outstanding interest
1. Shendure J, Mitra RD, Varma C, Church GM: **Advanced sequencing technologies: methods and goals.** *Nat Rev Genet* 2004, **5**:335-344.
 2. Blazej RG, Kumaresan P, Mathies RA: **Microfabricated bioprocessor for integrated nanoliter-scale Sanger DNA sequencing.** *Proc Natl Acad Sci USA* 2006, **103**:7240-7245.
Pushing the limits of Sanger sequencing. Amplifying, purifying and sequencing DNA in a single laboratory-on-a-chip device.
 3. Zhang K, Martiny AC, Reppas NB, Barry KW, Malek J, Chisholm SW, Church GM: **Sequencing genomes from single cells by polymerase cloning.** *Nat Biotechnol* 2006, **24**:680-686.
Sequencing of bacterial genomes after multiple displacement amplification (MDA) of chromosomal DNA from single cells with ϕ 29 DNA polymerase.
 4. Shendure J, Porreca GJ, Reppas NB, Lin X, McCutcheon JP, Rosenbaum AM, Wang MD, Zhang K, Mitra RD, Church GM: **Accurate multiplex polony sequencing of an evolved bacterial genome.** *Science* 2005, **309**:1728-1732.
With [5**], this work represents efforts at the leading edge of non-electrophoretic sequence determination after massively parallel amplification of genomic DNA. Both groups amplified DNA by emulsion PCR and bead capture. Shendure *et al.* carried out sequencing by automated cycles of hybridization and ligation of fluorescent probes. Margulies *et al.* used a pyrosequencing protocol.
 5. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bembem LA, Berka J, Braverman MS, Chen YJ, Chen Z *et al.*: **Genome sequencing in microfabricated high-density picoliter reactors.** *Nature* 2005, **437**:376-380.
See annotation to [4**].
 6. Dressman D, Yan H, Traverso G, Kinzler KW, Vogelstein B: **Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations.** *Proc Natl Acad Sci USA* 2003, **100**:8817-8822.
 7. Ramanathan A, Pape L, Schwartz DC: **High-density polymerase-mediated incorporation of fluorochrome-labeled nucleotides.** *Anal Biochem* 2005, **337**:1-11.
 8. Lacenere C, Garg MK, Stoltz BM, Quake SR: **Effects of a modified dye-labeled nucleotide spacer arm on incorporation by thermophilic DNA polymerases.** *Nucleosides Nucleotides Nucleic Acids* 2006, **25**:9-15.
 9. Seo TS, Li Z, Ruparel H, Ju J: **Click chemistry to construct fluorescent oligonucleotides for DNA sequencing.** *J Org Chem* 2003, **68**:609-612.
 10. Meng Q, Kim DH, Bai X, Bi L, Turro NJ, Ju J: **Design and synthesis of a photocleavable fluorescent nucleotide 3'-O-allyl-dGTP-PC-Bodipy-FL-510 as a reversible terminator for DNA sequencing by synthesis.** *J Org Chem* 2006, **71**:3248-3252.
 11. Bi L, Kim DH, Ju J: **Design and synthesis of a chemically cleavable fluorescent nucleotide, 3'-O-allyl-dGTP-allyl-bodipy-FL-510, as a reversible terminator for DNA sequencing by synthesis.** *J Am Chem Soc* 2006, **128**:2542-2543.
 12. Sakata T, Miyahara Y: **DNA sequencing based on intrinsic molecular charges.** *Angew Chem Int Ed Engl* 2006, **45**:2225-2228.
Enzymatic DNA extension detected when templates are immobilized on the surface of a field-effect transistor, based on the additional negative charge of the extended strand, rather than fluorescence. It is unclear how long reads will be achieved.
 13. Drmanac R, Drmanac S, Chui G, Diaz R, Hou A, Jin H, Jin P, Kwon S, Lacy S, Moer B *et al.*: **Sequencing by hybridization (SBH): advantages, achievements, and opportunities.** *Adv Biochem Eng Biotechnol* 2002, **77**:75-101.
 14. Edwards JR, Ruparel H, Ju J: **Mass-spectrometry DNA sequencing.** *Mutat Res* 2005, **573**:3-12.
 15. Stephan J, Dorre K, Brakmann S, Winkler T, Wetzel T, Lapczynska M, Stuke M, Angerer B, Ankenbauer W, Foldes-Papp Z *et al.*: **Towards a general procedure for sequencing single DNA molecules.** *J Biotechnol* 2001, **86**:255-267.
 16. Braslavsky I, Hebert B, Kartalov E, Quake SR: **Sequence information can be obtained from single DNA molecules.** *Proc Natl Acad Sci USA* 2003, **100**:3960-3964.
 17. Levene MJ, Korch J, Turner SW, Foquet M, Craighead HG, Webb WW: **Zero-mode waveguides for single-molecule analysis at high concentrations.** *Science* 2003, **299**:682-686.
 18. Church G, Deamer DW, Branton D, Baldarelli R, Kasianowicz J: **Characterization of individual polymer molecules based on monomer-interface interactions.** US Patent 1998, 5 795 782.
 19. Kasianowicz JJ, Brandin E, Branton D, Deamer DW: **Characterization of individual polynucleotide molecules using a membrane channel.** *Proc Natl Acad Sci USA* 1996, **93**:13770-13773.
 20. Akeson M, Branton D, Kasianowicz JJ, Brandin E, Deamer DW: **Microsecond time-scale discrimination among polycytidylic acid, polyadenylic acid and polyuridylic acid as homopolymers or as segments within single RNA molecules.** *Biophys J* 1999, **77**:3227-3233.
 21. Ashkenasy N, Sánchez-Quesada J, Bayley H, Ghadiri MR: **Recognizing a single base in an individual DNA strand: a step toward nanopore DNA sequencing.** *Angew Chem Int Ed Engl* 2005, **44**:1401-1404.
An important step towards DNA sequencing by the nanopore approach. Demonstrates both single base resolution and base recognition in an immobilized single-stranded DNA. Must be extended to a moving strand.
 22. Astier Y, Braha O, Bayley H: **Toward single molecule DNA sequencing: direct identification of ribonucleoside and deoxyribonucleoside 5'-monophosphates by using an engineered protein nanopore equipped with a molecular adapter.** *J Am Chem Soc* 2006, **128**:1705-1710.
Identification of all four nucleoside monophosphates by single-molecule detection with a protein nanopore. Again demonstrates base recognition, and may be a means for reagent-free base detection in exonuclease sequencing.
 23. Shapovalov G, Lester HA: **Gating transitions in bacterial ion channels measured at 3 microseconds resolution.** *J Gen Physiol* 2004, **124**:151-161.
 24. King GM, Golovchenko JA: **Probing nanotube-nanopore interactions.** *Phys Rev Lett* 2005, **95**:216103.
 25. Krieger K: **Nanopore arrays for DNA sequencing.** *Anal Chem* 2006, **78**:3486.
 26. Kang X, Gu L-Q, Cheley S, Bayley H: **Single protein pores containing molecular adapters at high temperatures.** *Angew Chem Int Ed Engl* 2005, **44**:1495-1499.
Protein pores are stable at the high temperatures that might be required nanopore sequencing.
 27. Shenoy DK, Barger WR, Singh A, Panchal RG, Misakian M, Stanford VM, Kasianowicz JJ: **Functional reconstitution of protein ion channels into planar polymerizable phospholipid membranes.** *Nano Lett* 2005, **6**:1181-1185.
Protein pores can function in polymerized bilayers, which are likely to be more stable than conventional lipid bilayers.
 28. Gufier PC, Pum D, Sleytr UB, Schuster B: **Highly robust lipid membranes on crystalline S-layer supports investigated by electrochemical impedance spectroscopy.** *Biochim Biophys Acta* 2004, **1661**:154-165.

29. Holden MA, Bayley H: **Direct introduction of single protein channels and pores into lipid bilayers.** *J Am Chem Soc* 2005, **127**:6502-6503.
30. Holden MA, Jayasinghe L, Daltrop O, Mason A, Bayley H: **Direct transfer of membrane proteins from bacteria to planar bilayers for rapid screening by single-channel recording.** *Nat Chem Biol* 2006, **2**:314-318.
- With [29], demonstrates means for the direct transfer of protein nanopores into lipid bilayers. Potential use for the fabrication of nanopore arrays.
31. Lagerqvist J, Zwolak M, Di Ventra M: **Fast DNA sequencing via transverse electronic transport.** *Nano Lett* 2006, **6**:779-782.
32. Gracheva ME, Xiong A, Aksimentiev A, Schulten K, Timp G, Leburton J-P: **Simulation of the electric response of DNA translocation through a semiconductor nanopore-capacitor.** *Nanotechnology* 2006, **17**:622-633.
33. Mitra RD, Butty VL, Shendure J, Williams BR, Housman DE, Church GM: **Digital genotyping and haplotyping with polymerase colonies.** *Proc Natl Acad Sci USA* 2003, **100**:5926-5931.
34. Henrickson SE, Misakian M, Robertson B, Kasianowicz JJ: **Driven DNA transport into an asymmetric nanometer-scale pore.** *Phys Rev Lett* 2000, **85**:3057-3060.
35. Bates M, Burns M, Meller A: **Dynamics of DNA molecules in a membrane channel probed by active control techniques.** *Biophys J* 2003, **84**:2366-2372.
36. Bayley H, Cremer PS: **Stochastic sensors inspired by biology.** *Nature* 2001, **413**:226-230.
37. Howorka S, Cheley S, Bayley H: **Sequence-specific detection of individual DNA strands using engineered nanopores.** *Nat Biotechnol* 2001, **19**:636-639.