Paul Dear
heads a group at the MRC Laboratory of Molecular Biology in Cambridge, working on genome mapping and on the development of new approaches to genomics.

One by one: Single molecule tools for genomics

Paul H. Dear

Date received (in revised form): 14th November 2002

Abstract

Much of the effort in any genomics programme arises from the need to generate and purify large numbers of identical molecules, since most analytical tools rely on the analysis of bulk DNA. Biological steps such as bacterial cloning — commonly used to prepare bulk samples of defined DNA fragments — are capricious and introduce their own restrictions and distortions. The analysis of single molecules, either directly or by *in vitro* enzymatic amplification, makes possible the examination of native genomic DNA without the complications and restrictions of biological propagation. Techniques already exist for the *in vitro* propagation of genomic fragments and for genome mapping, and offer the advantages of speed, flexibility and predictable behaviour. Single molecule sequencing, for which many approaches are being developed, is more challenging, but offers even greater rewards in terms of throughput and read length.

Keywords: *single molecule techniques, cell-free cloning, DNA sequencing, single molecule sequencing, genome mapping*

INTRODUCTION

The Sanger Institute, Genoscope and Stanford are some of the world's largest genome centres. But, between them, they have not yet sequenced a single molecule of DNA. True, they have sequenced tens of millions of cloned fragments and a good many polymerase chain reaction (PCR) products — but a single molecule of native genomic DNA? Not one has yet been sequenced.

And therein lies the problem: all of the conventional tools of genomics deal not with single DNA molecules but with billion upon billion of identical copies. Generating and purifying these herds of fragments — by cloning, subcloning, PCR, plasmid preps and electrophoresis — represents much of the effort of any genomics programme.

This need to analyse DNA in bulk not only makes life complicated, it also imposes severe limitations on what can be done. Cloning, for instance, sets the upper limit on the size of fragments which can easily be analysed and, as this limit is approached, artefacts and biases caused by cloning become more frequent and acute. Even small fragments, cloned in bacterial cells, are prone to such artefacts, or may interfere with the host's functioning so greatly as to be effectively uncloneable. Such problems are the reason why the initial 'shotgun phase' of a sequencing project (in which randomly chosen cloned fragments are sequenced to find those which overlap) is often only the beginning of a lengthy process to resolve errors and close recalcitrant gaps. And, when it comes to the sequencing itself, there are sound theoretical reasons (discussed below) which make it most unlikely that it will ever be possible to read more than a few thousand bases from 'bulk' DNA templates.

We are so used to working with bulk DNA that we take it for granted. When the assembly instructions for a flat-pack wardrobe show one door handle being fitted to one door, you hope to find approximately one of each in the box. But if your cloning kit contained only the single vector molecule depicted on the protocol sheet, you'd be surprised. 'A piece of DNA' always means 'lots of pieces just like this one'.

Life would be far simpler if we could look at native genomic DNA in the same way that the cell does — one molecule at a time. This would enable us to work

Paul H. Dear,
MRC Laboratory of Molecular Biology,
Hills Road,
Cambridge,
CB2 2QH UK

Tel: +44 (0)1223 402190
Fax: +44 (0)1223 412178
E-mail: phd@mrc-lmb.cam.ac.uk

directly from the genome, mapping and sequencing it without the convoluted steps needed to duplicate and purify selected pieces for bulk analysis. Over the last few years, this has begun to become possible. A new genomics tool kit is being assembled, allowing us to look at individual DNA molecules and to bypass the capricious steps of biological duplication. Some of the techniques in this tool kit make it possible to examine a single DNA molecule directly. Others ultimately rely on the analysis of bulk DNA, but generate it from single molecules by relatively predictable *in vitro* amplification rather than by *in vivo* propagation.

This paper focuses on single molecule alternatives to the three cornerstones of experimental genomics — cloning, mapping and sequencing — as more efficient tools for genome analysis. Many equally fascinating areas of research whose main aim is to understand the behaviour of single molecules as an end in itself have been deliberately avoided.

CLONING WITHOUT CLONING

Cloning is a biological method, and introduces biological distortions

As a method of duplicating a fragment of DNA, cloning is so taken for granted that it is worth pausing to consider just how strange it really is. A segment of DNA is spliced into a vector molecule and inserted into a living cell which, it is hoped, accepts the foreign molecule as a passenger and replicates it alongside its own genome. The cell is cultured and then its descendants harvested and the recombinant molecules purified. Inevitably, not all cloned fragments are inert and passive in the host cells — the whole point of DNA is to be biologically active — which leads to the aforementioned problems of clone artefacts and biases.

In principle, it should be possible to replace biological cloning with *in vitro* amplification, replicating single molecules as straightforwardly as we photocopy a sheet of paper.

Molecular cloning of RNA

The first demonstration of such an approach was performed not with DNA, but with RNA. It had long been known that replicase of the phage Qβ[1] was capable of exponentially amplifying certain RNA templates in an isothermal reaction.[2] These templates — known as RQ-RNA (*R*eplicable by *Q*β polymerase) — share secondary structural motifs which resemble the natural Qβ phage RNA and which appear to prime the replication reaction. As early as 1968, attempts were made to 'clone' RQ-RNA by *in vitro* amplification of single molecule template dilutions using Qβ replicase,[3] although the results were later called into question by the discovery of apparently 'spontaneous' generation of RNA in this system, even in the absence of added template.

Only after the pursuit of many red herrings was it demonstrated that the 'spontaneous generation' of RNA by Qβ replicase was due to the amplification of RQ-RNA molecules originating as contaminants. Efficient replication of single molecules of both contaminant and intentionally added RQ-RNAs was demonstrated in agarose layers containing the replicase and necessary reagents: each template molecule nucleated a 'molecular colony' of amplification products as a halo containing several nanograms of RNA.[4] This work was later extended and the method improved, by keeping the enzyme in an agarose film, and introducing the ribonucleotide triphosphates and RQ-RNA template molecules in an overlaid sheet of nylon membrane.[5] Each template molecule on the membrane gave rise to a distinct spot of amplification products — a true 'molecular clone'.

Attempts to exploit this system as a general means of *in vitro* RNA cloning were made, using naturally occurring RQ-RNA molecules as vectors.[6,7] Unfortunately, it was found that, with some possible exceptions,[8] the insertion of significant lengths of heterologous RNA into RQ-RNA abolished its ability to replicate efficiently *in vitro*. It seems that the main cause of this inhibition is the formation of double-stranded RNA by the two complementary strands generated during replication, disrupting the intramolecular secondary structure necessary for further replication.[7,9] Nevertheless, the presence of a coupled translation system (either *in vitro* or *in vivo*) alleviates this inhibition, presumably by sequestering the coding RNA strands into ribosomes and leaving the non-coding strands free for further replication.[9] This system has already been shown to augment the efficiency of continuous-flow *in vitro* translation systems, in which recombinant RQ-messenger RNA (mRNA) molecules are continuously regenerated.[10]

Expression libraries from single mRNA molecules

It has been proposed[9] that a system for cell-free gene cloning and expression could be developed, in which recombinant RQ-mRNA molecules are dispersed in an immobilised medium supporting both Qβ replication and translation. Each recombinant molecule would yield a 'molecular colony' containing both the replicated RQ-mRNA and its translation products, accessible for screening or selection. The potential for such a system in both proteomics and protein engineering is considerable, particularly as the possible toxicity of the proteins to living cells is irrelevant.

Replicating single molecules of DNA — polonies and limiting dilution PCR

In their papers demonstrating RNA cloning with the Qβ system, the authors suggested that other enzymatic amplification reactions could be applied for the analogous propagation of DNA.[5,9]

Single molecule PCR can replace bacterial cloning

PCR[11] was well known by this time and had been found capable of amplifying single template molecules by large amounts.[12] In principle, PCR could be used to generate molecular clones of single DNA templates, provided that such single molecules could be isolated and furnished with suitable priming sites.

A protocol for achieving this was demonstrated in 1999 by Mitra and Church,[13] building on the Qβ system of Chetverin's group. They dispersed DNA fragments in an acrylamide film bonded to a glass slide, at dilutions great enough to ensure that most molecules were well separated from their neighbours. (In this demonstration, the DNA fragments were synthetic constructs with known terminal sequences.) The acrylamide film contained *Taq* polymerase, primers complementary to the template termini and the necessary reagents for PCR. Thermocycling of the film led to amplification of the entrapped template molecules, producing molecular clones ('PCR colonies' or 'polonies') which appeared as discrete spots of fluorescence when stained with the DNA-binding dye SyBr Green I. Diffusion, which would tend to disperse the amplification products during cycling, was limited by immobilising one of the PCR primers. They also demonstrated that the films of polonies could be replicated by a form of 'contact printing' in which amplified template was transferred directly to a second acrylamide film for re-amplification (Figure 1).

In this way, the authors argued, a complete cell-free cloning system could be envisaged. Genomic DNA fragments, ligated into a suitable DNA 'cassette' carrying known priming sites, would be dispersed in acrylamide films, and up to 5 million distinct molecules amplified as discrete polonies on a single microscope slide — a high-density, cell-free version of a bacterial clone plate.

Such a system may well be an ideal means of propagating DNA fragments without the biases of cellular cloning, but how could it be used in practice? Conventional clone plates are harvested by using a toothpick or robotic probe to transfer a sample of each colony to a microtitre well for further growth. The inventors of the polony system demonstrated the analogous step —

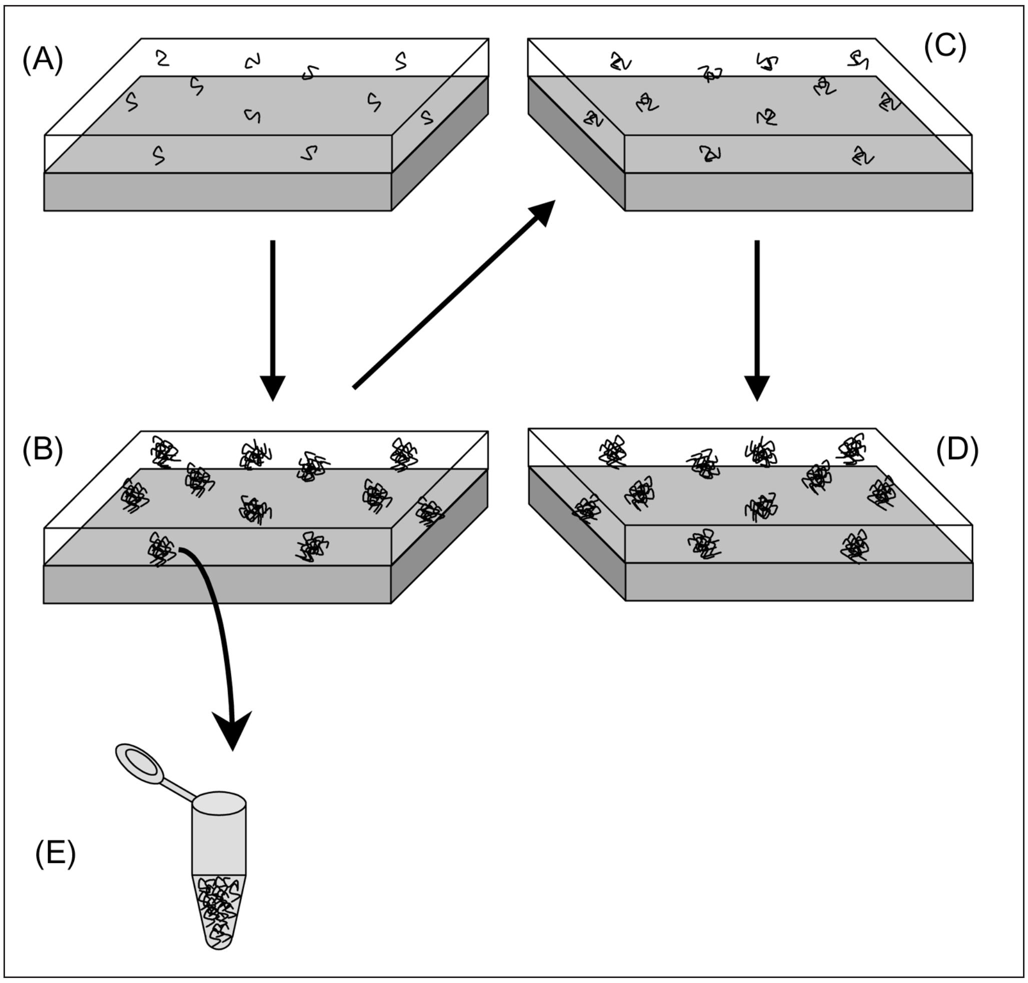

Figure 1: Molecular cloning by PCR. (A) DNA molecules (short irregular lines) are dispersed in a gel bonded to a solid support (shaded) and amplified *in situ* by thermocycling. Each molecule gives rise to a cluster of replicas — a PCR colony or 'polony'. (B) The polony slide can be replicated as a mirror image (C) which can be similarly amplified (D). Alternatively, single polonies can be toothpicked out of the gel and re-amplified in a conventional liquid PCR

Molecular clones can be harvested like bacterial clones, or sequenced *in situ*

toothpicking a single polony into a conventional liquid PCR reaction, where it could be amplified to yield sufficient product for visualisation by gel electrophoresis or, in principle, for sequencing.[13] This demonstrates that the polony system could perhaps provide the ideal means of obtaining relatively unbiased fragments for shotgun sequencing projects, either replacing or supplementing conventional small-insert bacterial clones. The cost of PCR amplification of each 'clone' may be a disincentive (especially when compared with the cost of growing and processing conventional clones), but this is likely to become less significant as reagent costs decline and as more efficient sequencing machines require less and less template.

There is another route by which polonies can be exploited as sequencing templates. The authors suggested in their original paper that pyrosequencing — a method in which a few tens of bases of sequence can be read from a template by monitoring the incorporation of nucleotides by a polymerase — could be used to read the terminal sequences of many thousands or millions of polonies *in situ*. Many other such 'mini-sequencing' methods have been proposed in various contexts and are valuable in, for example, comparing the sequences in a complementary DNA (cDNA) library (or, in this case, a polony slide made from cDNAs) with a set of reference sequences. A different approach — also capable of reading a few bases of sequence *in situ* — was reported by the inventors of the polony system.[13]

An alternative to the polony approach would be to isolate single templates in individual reaction chambers, before amplifying them in a conventional liquid-phase PCR. This can be done, without resorting to micromanipulation or sophisticated biophysics, by an approach which has long been used in the creation of monoclonal libraries and other traditional cloning methods: limiting dilution. If either cells or (in this case) DNA molecules are diluted sufficiently, the solution can be dispensed into a microtitre plate such that each well, on average, contains only a single DNA molecule. Of course, some wells will contain two or more molecules, and some will contain none, but the Poisson distribution dictates that 37 per cent of wells will contain just one molecule.

Protein libraries from single DNA molecules

This approach has been used, for example, by Nakano and colleagues, who dispersed mutagenised DNA sequences encoding either green fluorescent protein[14] or single-chain antibodies[15] among the wells of microtitre plates at great dilution, amplified them using a two-step PCR protocol, and then used products for *in vitro* transcription and translation. In each case, they clearly demonstrated that a proportion of the wells had contained single DNA molecules which yielded a distinct protein product.

Molecular cloning could create unbiased libraries for sequencing

Naturally, such an approach (although without the transcription and translation) could be used to create shotgun libraries for sequencing, by limiting the dilution of genomic fragments ligated to linker sequences which serve as PCR priming sites. One of its main drawbacks is that, regardless of the degree of dilution used, only a minority of the wells will contain a single template molecule at the outset. Wells containing no templates (and hence no product) are obviously wasted, but could be readily identified before feeding into the sequencing process. Those which contained multiple template molecules will yield mixed amplification products, identifiable only by further analysis or by their becoming apparent when an attempt is made to sequence them. Nevertheless, such approaches may well be worth considering, particularly for those genomes which, having extreme base compositions or frequent unstable sequences, are difficult to clone in conventional cell-based systems.

Surely, though, all such cell-free cloning systems suffer from one overriding limitation: amplification errors? In fact, the situation is not as bad as one might expect, given the relatively high error rates of most polymerases *in vitro*.[16] An error in the first cycle of amplification will be present in only one of the four resulting DNA strands, and hence in only a quarter of the final product molecules. An error in the second round will, likewise, affect only one-eighth of the products; in the third round, a sixteenth; and so on. In other words, only errors arising in the first three or four cycles of amplification will affect a large enough proportion of the final product to have any impact when this is used in a sequencing reaction. Other errors, although numerous, are each present in such a tiny proportion of the molecules as to be invisible if the amplified DNA is sequenced. The frequency of such early-cycle errors when using *Taq* polymerase will be on the order of 0.03 per cent per nucleotide position — ie one base in every 3 kilobases (kb), three or four times the per-cycle error rate[16] — and higher-fidelity polymerases and enzyme cocktails[16–18] can reduce this even further (Figure 2).

Cell-free systems, then, are beginning offer an alternative to *in vivo* cloning for small nucleic acids, especially in the contexts of shotgun sequencing and protein expression. But can they compete with traditional methods when it comes to cloning the larger fragments needed, for instance, in physical mapping? Here, the answer appears to be no, at least for the present. Neither Qβ-like systems nor PCR (with currently available enzymes) can reliably and routinely amplify templates much larger than a few kilobases, even when starting with large quantities of template DNA. Bacterial cloning, in contrast, can stably propagate

PCR errors are not as severe as one might assume

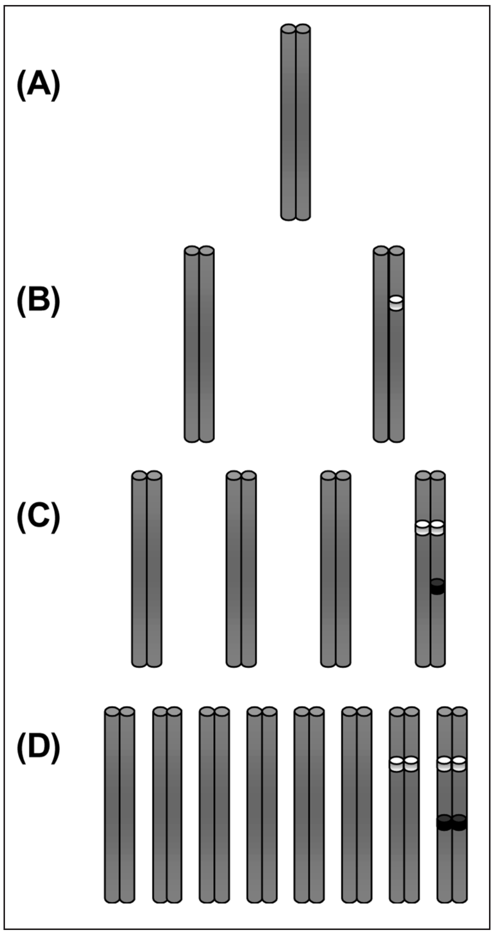

Figure 2: Propagation of errors during PCR. Starting with one double-stranded DNA molecule (A), an error arising during the first cycle of amplification (white segment) affects only one strand in the next generation of products (B), and only one-quarter of the molecules in subsequent rounds (C,D). An error arising during the second round (black segment) affects only one-eighth of the subsequent products. Errors in later rounds propagate to even smaller proportions of the final products of amplification

***In vitro* methods cannot match biological cloning for large molecules – yet**

many (although not all) fragments of up to several hundred kilobases using BAC (bacterial artificial chromosome) vectors.[19] Yeast vectors[20] can accommodate inserts of over a megabase, albeit with a high frequency of clone artefacts among the largest such clones.

Nevertheless, it will surely not be long before systems are developed which can reliably do *in vitro* what *Escherichia coli* can do on an agar plate, and replicate single DNA molecules of a respectable size.

MAPPING WITHOUT BIOLOGY

Genome mapping — finding the locations of defined markers or sequences within a genome — is often a prelude to complete genome sequencing. Dense maps provide a framework around which to assemble fragmentary shotgun sequence data. Maps are also valuable as resources in their own right: an accurate map makes it easier to isolate genes of interest, to home in on particular regions for sequencing or to compare the organisations of different species' genomes.

Most of the conventional methods for genome mapping rely on the analysis of bulk DNA, and have a major biological component. Physical mapping aims to find cloned fragments (usually BACs or other large-insert clones) which represent overlapping segments of the genome; the overlaps are determined either by two clones containing the same sequence-based marker,[21,22] or by their sharing several restriction fragments of the same sizes.[23] Genetic linkage mapping[24] follows the way in which polymorphic (that is, variable) sequence markers are passed from parents to offspring: markers that are adjacent in the genome tend to be inherited together, while remote markers are inherited independently as a consequence of meiotic shuffling.

Radiation hybrid mapping[25–27] is intermediate in its approach between clone-based mapping and genetic linkage: the genome is broken by irradiation to give fragments which are propagated as complex clones in eukaryotic host cells. Sequence markers which lie close together in the genome typically remain linked on the same fragment after irradiation, and hence tend to be found together in the recombinant ('radiation hybrid') cells.

Biological complexities limit conventional mapping methods

All of these methods are relatively indirect means of answering the simple question 'what is the order of these markers along the chromosomes?' Biological factors — cloning artefacts, the vagaries of meiosis and the biased retention of some sequences by radiation hybrid cells — limit both the applicability and the reliability of these approaches, so that two independent maps of the same genome seldom agree perfectly with one another.[28] Can single molecule methods, bypassing the biological steps, do any better?

Mapping by direct imaging — FISH and optical mapping

In fact, single molecule mapping methods have a respectable history, dating back to the 1950s when the human karyotype was first defined[29] and trisomy 21 characterised[30] by direct microscopic observation of chromosomes. (Genetic linkage mapping, by the way, also has a respectable pedigree: the order of several genetic factors along the chromosomes of *Drosophila* was inferred in 1913 by Sturtevant.[31])

FISH allows elegantly direct mapping on single DNA molecules

Direct analysis of chromosomes or of DNA has evolved by the use of DNA probes and *in situ* hybridisation. These probes can be hybridised to their complementary sequences in metaphase chromosomes fixed to glass slides, and visualised using either radioactive labelling or, nowadays, by fluorescent labels of various colours. Such fluorescence *in situ* hybridisation, or FISH,[32–34] can identify the location of a probed sequence relative to the characteristic banding pattern produced by staining the chromosomes with suitable dyes. If two or more probes, labelled with different fluorophores, are hybridised simultaneously, then their relative positions can be found. Analogous methods, when applied to interphase nuclei or extended chromatin fibres,[32,35–37] achieve higher resolution (down to a few kilobases, as opposed to a few megabases for conventional FISH), as the DNA is less heavily condensed.

FISH is a very elegant and direct means of mapping: no amount of genetic linkage data is quite as compelling as actually seeing paired fluorescent spots on the sister chromatids of a metaphase spread. It is, however, labour intensive and ill-suited to high throughput applications, requiring considerable expertise and yielding information on only one or a few markers at a time. More recently, another mapping method has been devised which relies on the direct observation of single DNA molecules, yielding rich information which integrates well with other genomic tools. This method, Optical Mapping,[38–40] was developed by David Schwartz and colleagues at New York University. DNA — from a bacterial genome, for example — is spread and immobilised on a modified glass surface in a way that ensures that the molecules are extended and nearly linear. A buffer solution and restriction enzyme are added, cleaving the DNA at the restriction sites. When the immobilised DNA is stained with a fluorescent dye and imaged microscopically, individual molecules appear as lines broken at the restriction sites. Schwartz, Mishra and colleagues have developed not only the experimental tools but the software necessary to extract and process information from the microscopic images, deriving fragment sizes from the observed contour lengths. Data from many molecules and several different restriction enzymes can be amalgamated, producing complex and accurate maps[41] of the restriction sites. These maps address one of the major requirements of genome sequencing — the validation of long-

Optical Mapping can provide restriction fingerprints to guide sequence assembly

range sequence assembly — by comparing the pattern of restriction sites predicted from the sequence with that seen by optical mapping.

An alternative approach to the same problem was demonstrated more recently by Schäfer's group.[42] They used complementary oligonucleotides to tether fluorescently stained single DNA molecules to polystyrene microspheres which could be held by optical tweezers. By causing a buffer solution to flow past the fixed molecule, and adding a restriction enzyme to this flow, they were able to observe the detachment of restriction fragments from the tethered DNA and to estimate their size from their degree of fluorescence. It is not clear, however, that such a system has any advantage over optical mapping in terms of simplicity or accuracy.

HAPPY mapping — co-segregation in single molecule samples

HAPPY mapping produces marker-based maps without biological complications

FISH and optical mapping represent single molecule mapping in its most direct form. A different approach is used by the author's own favourite technique: HAPPY mapping.[43,44] Here, bulk genomic DNA is prepared and broken randomly either by mechanical shearing or (where larger fragments are needed) by irradiation. This fragmented DNA is diluted greatly, and dispensed into about 100 samples which, on average, contain about one-half a genome's worth of DNA fragments. This set of samples — a 'mapping panel' — is then tested by PCR to find out which markers (short, unique sequences known to lie in the genome) are present in each member of the panel. Two markers which lie close together (compared with the average size of the fragments) will generally remain physically linked on the same DNA fragment after the initial breakage, and hence will tend to be found together in many of the samples in the mapping panel. Remote markers will lie on distinct fragments after DNA breakage, and hence occur independently of each other in various members of the mapping panel. By examining the frequency with which any two markers co-segregate in this way, their physical distance can be computed, and hence a map can be made (Figure 3).

HAPPY mapping, then, is closely analogous to radiation hybrid mapping, but relies on simple limiting dilution, rather than cloning in hybrid cells, as a means of sampling a portion of the genome; it is, in effect, radiation hybrid mapping without the radiation hybrids. This replacement of the biological step with a purely *in vitro* one gives significant advantages, and some limitations. The main advantages are the simplicity with which HAPPY panels can be made from any genome, its high resolution (if the DNA is broken finely at the outset) and its freedom from artefacts and distortions. Radiation hybrids do have one advantage over HAPPY mapping panels: the chromosomal fragments which are propagated in radiation hybrids can be larger (up to tens of megabases) than the longest DNA fragments (up to 2 or 3 megabases[45]) which can be diluted and dispensed into a HAPPY panel. This means that RH mapping is effective in making very coarse- to medium-resolution maps, whereas HAPPY mapping spans the range from medium[45,46] to very fine[47,48] resolution.

In summary, single molecule methods have some distinct advantages over bulk DNA approaches when it comes to genome mapping. Above all, they are more universally applicable than methods which depend on biological processes, and are less vulnerable to errors and artefacts.

SINGLE MOLECULE SEQUENCING

If the arguments for developing single molecule alternatives to cloning and mapping are strong, then those for developing single molecule sequencing are overwhelming. The ability to read the sequence of native fragments of genomic DNA would not only eliminate the need for subcloning and complex template

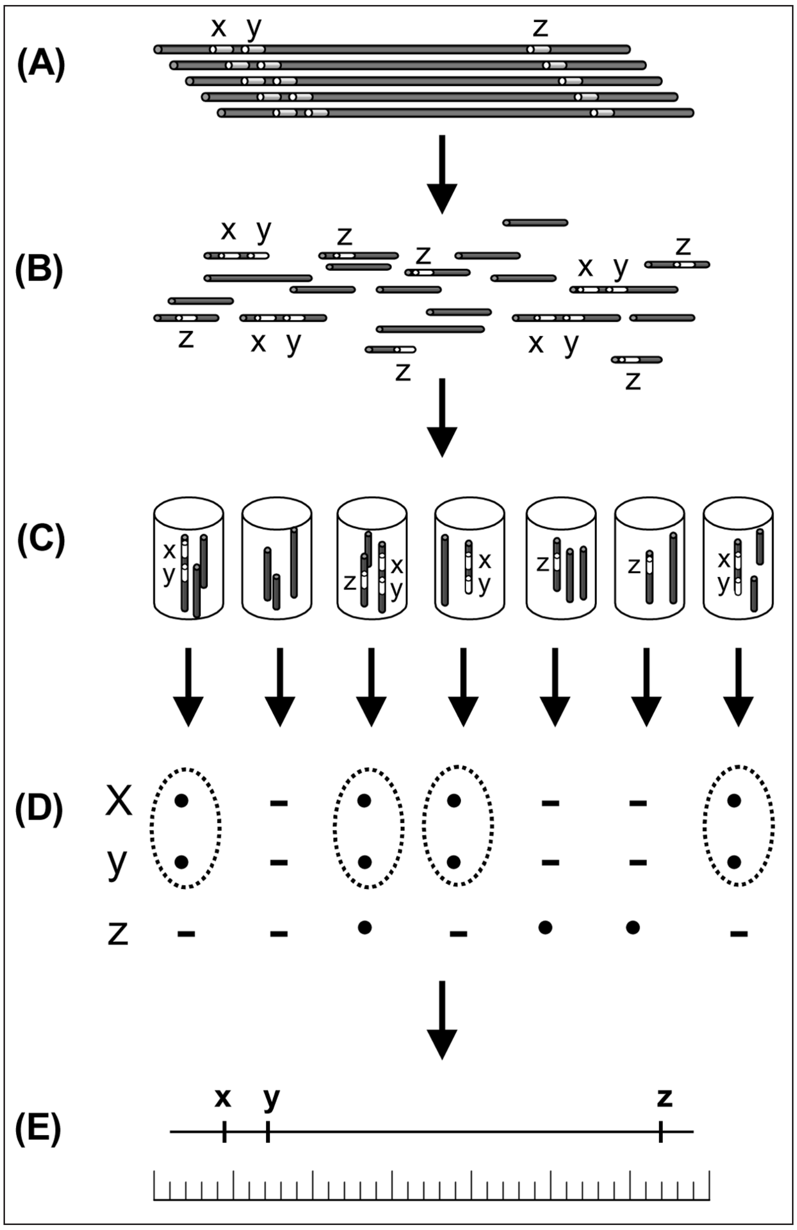

Figure 3: HAPPY mapping. The starting point is genomic DNA (A), carrying marker sequences to be mapped (white segments x, y, z). The DNA is broken into random fragments (B) and dispensed into a series of samples (C), each containing only a few molecules. Each sample is screened by single molecule PCR to determine which of the markers it contains (D); closely linked markers x and y tend to found to be found together (co-segregate) in the same samples (dotted ellipses). From the frequency of co-segregation, the distances between markers can be computed, giving a map (E)

Single-molecule sequencing must compete with improvements to conventional sequencing methods

preparation, but could potentially offer much higher rates of sequencing (either by rapid reading of molecules one at a time or by parallel reading of vast numbers of different templates) and, maybe, lower costs.

But, before exploring single molecule techniques, it is worth mentioning that many alternative sequencing technologies are being developed which, although not operating at the single molecule level, nevertheless promise great increase in throughput. Most productive in recent years has been the refinement of instruments[49–51] for reading conventional Sanger sequencing reactions (including microfluidic electrophoresis devices for greater speed and reduced reagent consumption[52,53]). Working on a very different principle, sequencing by hybridisation (SBH) determines which short sequences are present within a larger piece of DNA, by hybridisation to an array of oligonucleotides. Although the hope of *de novo* sequencing by this means[54] has not been realised, the spirit of SBH has been revived very successfully as a means of mutation detection and 're–

sequencing' or comparing the sequence of an individual with that of a known reference which is represented by arrayed oligonucleotides.[55,56]

Several non-single molecule methods can read sequence 'fingerprints' efficiently

Several methods have been developed for efficiently reading short 'sequence signatures' of a few tens of bases from each template, typically for the detection of mutation or the characterisation of a large population of DNA molecules (for example, cDNAs). All of these methods, like traditional Sanger sequencing,[57] use bulk DNA rather than single molecules as the template. Pyrosequencing[58–60] monitors the release of pyrophosphate as successive nucleotides are incorporated into primed templates by a polymerase. Massively parallel signature sequencing[61,62] (MPSS) attaches a unique sequence 'tag' to each of millions of distinct cloned fragments, and in turn captures many copies of each fragment on an individual bead. The captured fragments can then be sequenced by repeatedly cleaving the terminal few bases and reading the newly exposed termini by a hybridisation method. Finally, mass spectrometry has been developed as a tool for reading short sequence signatures or comparing sequences[63,64] by measuring the masses of successively longer terminal fragments of DNA — a sort of gel-free analogue of the electrophoretic ladder produced in Sanger sequencing.

Only single-molecule methods promise very long sequence reads

Single molecule sequencing offers one possibility which no other technology is ever likely to provide: the ability to read very long DNA molecules. Whenever sequence information is read from bulk DNA, some means has to be found of synchronising the results from each molecule. In conventional Sanger sequencing (which offers the longest read length at the moment, up to a kilobase or so), each of many template molecules is replicated by a polymerase, terminating at a defined base (A,C, G or T) in the sequence. Gel electrophoresis provides the synchronising mechanism, sorting the truncated strands by size so that the sequence ladder can be read. Although both electrophoresis and enzyme systems have been refined immensely over the last decades, no such system is ever likely to reliably resolve a strand of, say, 10,000 nucleotides from one of 10,001, so electrophoretic resolution will always set an upper limit on the length of readable sequence. In pyrosequencing, the stepwise extension of the many template molecules eventually drifts out of phase, so that the sequencing signal becomes incoherent after a few tens of bases. Similar constraints apply to all other multi-molecule methods. Only if a single DNA molecule is read does the synchronisation issue disappear.

Why is it important to be able to read very long sequences? Well, quite simply because genomes are big.[65] Reassembling a large genome sequence — or even a 100 kb BAC clone — from thousands or millions of sub-kilobase reads is fraught with complications, like reassembling a shredded telephone directory. Larger reads reduce the effort disproportionately: if the read length is doubled, far less than one-half as many such reads are needed for a good assembly. Long reads also make it easier to span the conserved repeat sequences present in most genomes, and which can otherwise confound assembly. At present, sequencing a 100 kb bacterial clone involves subcloning, shotgun sequencing to perhaps six- or eight-fold redundancy, assembling and editing, and then a finishing phase to close the remaining gaps and resolve ambiguities — all of which require several man-days of effort even in a well-run genome centre. If such a sequence could simply be read, end to end, in one go, the effort saved would be enormous.

So, what approaches are being developed to enable true single molecule sequencing?

Scanning probe microscopy — DNA sequencing by Braille

As the physicist Richard Feynman pointed out,[66] it should be very easy to read the sequence of DNA: 'you just *look at the thing*!' He envisaged using electron microscopy to image directly the base pairs, so that the genetic information could be read like any other text.

Scanning probe microscopy can image single DNA molecules

Over the last 16 years, a range of related methods have been developed, which allow a probe — typically on the nanometre scale — to be scanned across the surface of a specimen. These techniques, collectively referred to as scanning probe microscopy (SPM), can be used to examine surfaces with resolutions of nanometres or below. Scanning force microscopy (SFM) measures the minuscule forces generated as the probe tip approaches the scanned surface, and vertical control of the probe allows a contour map to be created. (A related method, scanning tunnelling microscopy or STM, measures minuscule currents between the probe and sample rather than direct forces, but can likewise be used to map the contours of a surface; for an excellent review of SPM methods and applications, see ref. 67).

Direct imaging of the sequence of bases is hampered by the flexibility of DNA

As long ago as 1995, SFM was used successfully to image DNA, strikingly revealing the contours of the double helix.[68] Could such a method, perhaps applied to single-stranded DNA, be used to 'read' directly the sequence of DNA by identifying individual bases? Such a goal has not been met, largely because DNA, especially single-stranded DNA, does not lie like a rigid molecular model on a tabletop, but suffers distortions and twists which obscure the neat regiments of base pairs.

Scanning probe microscopy can image labels attached to single DNA molecules

Nevertheless, the ability to image DNA molecules on a sub-nanometre scale has found a number of applications related to sequencing. In particular, specific sequence features can be identified by SFM if they are first tagged with a bulky marker. In this way, the binding of sequence-specific proteins to their cognate sequences can be observed (see, for example, ref. 69). Single-base mismatches between two DNA strands have been imaged, by first treating the DNA with MutS, a protein which recognises the local disruption in the double helix.[70] A more general technique has been developed by Woolley *et al*.[71] in which sequence-specific oligonucleotides, tagged with bulky molecules, are allowed to anneal to complementary sequences in denatured DNA. The tags can be imaged by SFM and, critically for their application, different types of tag can be distinguished by their sizes. Since the oligos anneal to the DNA with single base specificity, they were able to use this method to characterise sequence variations at two distinct loci within a 10 kb template, and thereby determine the haplotype of individual molecules — representing a nano-scale analogue of FISH.

But these sensitive force-based techniques can be used to obtain another type of sequence-related data as well. Techniques related to SFM can be used to apply forces in the picoNewton range to DNA molecules, and to measure the resulting deflections. Why is this significant? Because the very forces which hold double-stranded DNA together — the interactions between paired bases — are precisely within this range. Rief *et al*.[72] used an SFM probe to stretch and relax hairpins repeatedly in DNA strands of various base compositions, unwinding and reforming regions of double helix. In this way, they were able to measure the forces which hold individual A–T and G–C base pairs together — at about 20 and 9 picoNewtons, respectively. (For reference, an apple sits on your hand with a force of about one Newton.)

Now, if it is possible to measure the base-pairing forces between two DNA strands and, if these forces differ for A–T and G–C bases, then surely it should be possible to use this approach to get some kind of sequence information from DNA? Essavez-Roulet *et al*.[73] demonstrated exactly this, using an elegant method which differs slightly from the SFM approach. They progressively peeled apart the two strands of a lambda DNA molecule tethered between a fine glass needle and a flat surface, and monitored the deflection of the needle to reveal the forces involved. The *average* force during this 'unzipping' was around 13

Peeling apart the two strands of a DNA molecule can directly reveal its base composition

picoNewtons — ie somewhere between Rief's measured strengths for A–T and G–C base pairs.[72] But, much more strikingly, the unzipping force fluctuated between about 11 and 14 picoNewtons along the length of the molecule and this fluctuation corresponded perfectly with the relative proportions of A–T and G–C basepairs along the known sequence of lambda DNA. Thus, they had measured the base composition profile of a large DNA molecule directly. More recently, the same group has used optical tweezers, coupled with an improved DNA configuration, to measure unzipping forces with far higher spatial and temporal resolution, reflecting sequence features as small as ten bases.[74] This is a long way short of obtaining sequence data, but even a coarse base composition profile of a long DNA molecule could be used as a guide to detect errors in sequence assemblies produced by shotgun methods (Figure 4).

Nanopores — tickertape readers for DNA

Nanopores could scan DNA molecules

If DNA cannot be read directly by SFM, then nanopore technology offers the next most direct means of sequencing.[75] The basic idea is elegantly simple: a voltage is applied across a nanometre-scale pore in an otherwise insulating membrane, causing a measurable current to flow through the pore. If a single strand of DNA is then threaded through the pore, each passing base will transiently occlude the pore and restrict the flow of current. Given the different shapes of the four bases, one might hope that the degree of occlusion would depend on the base and that the sequence could be read off as a series of current dips of different magnitudes.

In practice, some very formidable hurdles (or at least nanometre-scale ones!) have to be overcome before this becomes a practicable means of sequencing DNA. Most experiments to date have made use of the pores formed by alpha-haemolysin in lipid bilayers.[76] These pores are about 1.5 nm across at their narrowest point — comparable in scale to single-stranded DNA. It has been successfully shown that single-stranded nucleic acid molecules cause a measurable reduction in the ionic current flowing through such a pore.[77] More excitingly, the duration and extent of current reduction are measurably different for homopolymers of different nucleotides and different lengths, and tracts of purines and pyrimidines in a synthetic construct could be distinguished as they passed through the pore (Figure 5).[78]

The overwhelming problem with alpha-haemolysin pores is that the narrow

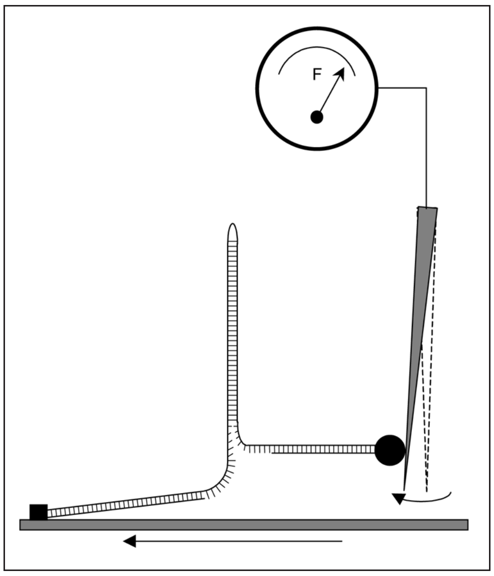

Figure 4: Measuring the unzipping force of DNA. A stem-and-loop DNA construct was tethered between a microscope slide (horizontal grey bar) and a fine glass microneedle (on right). As the microscope slide was moved leftwards, the deflection of the microneedle was recorded, revealing the force (F) needed to break successive base pairs as the vertical stem of the DNA was peeled apart (based on ref. 73)

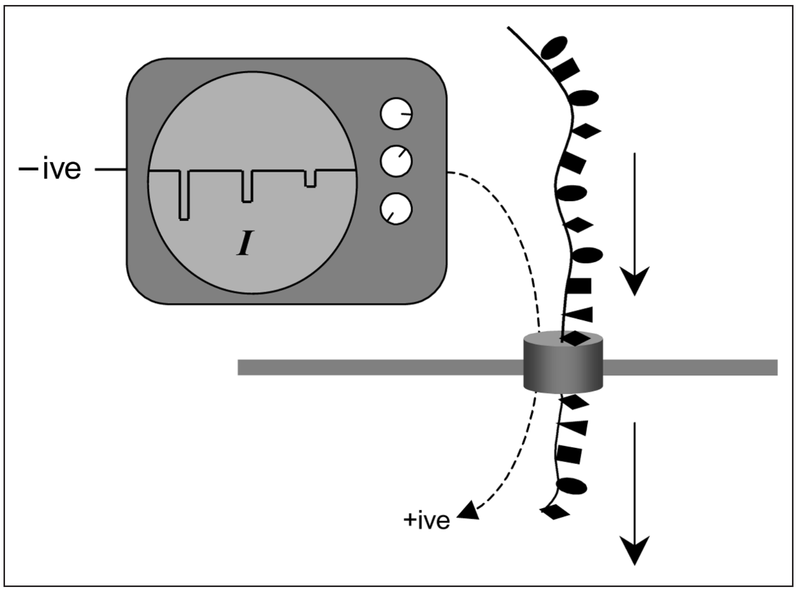

Figure 5: An idealised nanopore sequencing system. Single-stranded DNA (wavy line) is pulled through a pore (cylinder) in an impermeable membrane by an electric field. As each base (ellipses, rectangles, triangles, diamonds) passes through the pore, it briefly blocks the passage of ions (dotted line) to a greater or lesser degree, transiently reducing the ion current through the pore. The degree of current blockage reflects the bulk (and hence the identity) of each base as it passes through

Existing nanopores cannot discriminate single nucleotides

part of the channel is several nanometres long, so that several nucleotides occupy it at any one time. Therefore, the degree of current reduction at any instant is dependent not on a single base, but on six or eight consecutive bases. Moreover, the contribution of any one base to the current reduction is so small, and the passage of the DNA through the pore so rapid, that near-impossibly sensitive measurements would be needed to discern the contribution of a single base.[75]

Secondary structures or polymerases could control the DNA's passage through a nanopore

Nevertheless, the approach itself is sound, at least if DNA can be passed slowly enough through a channel which is both narrow and short. Some work has already shown that movement through alpha-haemolysin pores can be slowed by secondary structures such as hairpins in the DNA.[79] A similar braking effect might be exploited when sequencing naturally occurring double-stranded DNA, which would have to unwind progressively to allow one strand to pass through the pore. Alternatively, perhaps, a protein pore could be engineered with a DNA or RNA polymerase moiety built in at its mouth; in this case, either the template strand or the newly synthesized strand would feed through the pore at a rate limited by the polymerase.

In the meantime, alpha-haemolysin nanopores still have some potential applications, in detecting hybridisation which forms secondary structures either within or between DNA molecules. Vercoutere and colleagues[79] examined the effect of hairpin-shaped oligonucleotide structures on the current through such nanopores. A molecule with a double-stranded portion is unable to pass through the pore, and instead sticks in its mouth, reducing the ion current by about one-half. Only when the stem of the hairpin 'unzips' can the now single-stranded molecule pass through the pore, transiently blocking the current almost completely as it passes through the narrow throat of the channel. Most excitingly, the time taken before the hairpins unzip, and the degree to which they restrict the current in the meantime, are sensitively dependent on both the length of the base-paired stem and the presence of any mismatched bases within it, offering clear discrimination of these features down to the base-pair level.

Howorka's group exploit nanopores in a slightly different (and potentially more versatile) manner.[80,81] They first covalently tether an oligonucleotide to the lip of the pore, so that it dangles into the mouth of the channel but is not wide enough to block it appreciably. If

Nanopores can detect single-base mismatches

oligonucleotides which are complementary to the tethered DNA are presented to the 'mouth' side of the channel, they anneal briefly to the tethered oligo, forming a double-stranded molecule within the pore's mouth, and reducing the current much like Vercoutere's hairpins. As the annealed oligo dissociates after a few milliseconds, the freed strand zips though the narrow channel, causing a brief, almost complete abolition of current as it passes through. Oligos which are not complementary to the tethered DNA, in contrast, do not pause to anneal and only the transient complete reductions in current are seen as they pass unimpeded through the system.

An array of nanopores could identify polymorphisms in DNA

Using this system, Howorka's group have first measured the kinetics of duplex formation and dissociation within the mouth of the pore.[80] In their second paper,[81] they demonstrate that these nanopore/tethered oligo complexes can reliably detect single-base mismatches between the tethered oligo and the one presented in solution. This clearly has tremendous potential. They point out that an array of nanopores, each carrying a different tethered oligo, could be used to detect single nucleotide polymorphisms (SNPs) rapidly and efficiently, and they demonstrate this in a model system. In theory, a similar system could be used for DNA sequencing (operating on the same principle as sequencing by hybridisation), although constructing a suitably large array of nanopores with the requisite tethered oligos would probably not be feasible.

Sequencing by synthesis — watching DNA polymerase at work

Several approaches aim to watch over the shoulder of a single DNA polymerase at work

Both SFM and nanopore technology aim to read DNA in a more or less literal sense, much as we would read a book. Other approaches are less direct, including several that come under the heading 'sequencing by synthesis'. The unifying theme here is that the activity of the DNA polymerase is monitored as it extends a primed single-stranded template. Both pyrosequencing and several other systems fall under this heading, but these have not been applied to the sequencing of single molecules and (because of the problem of keeping multiple molecules 'in phase', as discussed above), they are not capable of reading more than a few tens of bases.

A true single molecule 'sequencing by synthesis' approach is being developed in a corporate setting, by the Cambridgeshire-based company Solexa.[82] In their approach, single-stranded DNA molecules are primed for extension and dispersed on a solid substrate. In the presence of a DNA polymerase, each template is extended by one base, by incorporating a nucleotide labelled with one of four different fluorophores for each of A,C, G and T. A sensitive optical system then images each molecule, identifying and recording the specific base incorporated into each template. The fluorescence is then abolished, and the process repeated. The aim is to be able to perform 20–30 such cycles, thereby reading 20–30 bases of sequence from each template molecule. This sounds interesting but not impressive, until one realises that up to 1 billion template molecules could be analysed simultaneously in this way on a 3 × 3 cm surface, making even MPSS look less massively parallel in comparison. The short read lengths make it unlikely that such a system will contribute towards genome sequencing, but, like MPSS, it would be ideally suited to re-sequencing, mutation detection or surveying complex DNA mixtures.

Other attempts at sequencing by synthesis aim to observe the activity of a single polymerase in real time, monitoring the incorporation of successive fluorescently labelled nucleotides by highly sensitive imaging of a tightly confined volume. Groups at both Caltech[83] and Cornell[84,85] are developing such systems. These approaches, requiring close monitoring of a single polymerase/ template complex, do not promise the massive parallelism of Solexa's strategy but may offer longer read lengths, if the polymerase can be monitored over a sufficient number of steps.

Sequencing by degradation — slicing DNA base by base

The sequencing by synthesis approaches described above are limited in their read length ultimately by the processivity of the polymerase (although in practice other factors come into play long, long before this limit is reached). It is far easier in general to take something apart than to put it together, and this is the tactic employed in sequencing by degradation.

Sequencing by degradation is an elegantly simple concept promising fast, long reads

As with most of these methods, the basic idea is attractively simple.[86] A single DNA molecule is tethered to a fixed support, and an exonuclease cleaves successive nucleotides from one end. These are carried away, either in a flow stream or electrophoretically, and detected and identified downstream. Given the high rates of exonuclease cleavage, and the fact that such a reaction can continue almost indefinitely, the read length of such a system should be limited only by the size of DNA that can be isolated and tethered — perhaps a hundred kilobases or more (Figure 6).

Single molecules of natural nucleotides are almost impossible to detect

Nevertheless, the technical hurdles which must be overcome are acute — perhaps more so than in the other single molecule sequencing strategies. First, there is the problem that single molecules of natural nucleotides are only very weakly fluorescent and therefore virtually impossible to detect. Therefore, it is necessary to first synthesise (by replication with polymerases) DNA templates which incorporate fluorescently labelled nucleotides in place of the natural ones.[87]

Great effort has been put into finding suitable fluorescent base analogues which can be incorporated efficiently into DNA by polymerases (see, for example, refs. 88–91), but the results to date are a long way short of perfection. Efficient substitution of all four bases with fluorescent analogues has been demonstrated for short molecules,[88] or substitution of two of the four bases for molecules up to 2.5 kb using a modified polymerase,[92] but complete substitution of very long DNA has not yet been possible. Indeed, given the difficulty of *in vitro* replication of very long DNA molecules in general (even using natural nucleotides), it is likely to be some time before fluorescent templates can be made which are long enough to realise one of the goals of single molecule sequencing — read lengths in excess of a few kilobases. It should be noted that substitution of all four bases is not essential: if any two bases can be fully substituted, then a complete sequence can be reconstructed from the combined reads obtained with different permutations.

If a suitably labelled template can be prepared, then it must be tethered in a system where the exonuclease can act and the nucleotides can be carried away (by either hydrodynamic or electrokinetic forces). A range of experimental set-ups have been tested, and it appears that these problems, at least, are soluble.

The greatest challenge remains the detection and identification of the single base molecules, even if these are

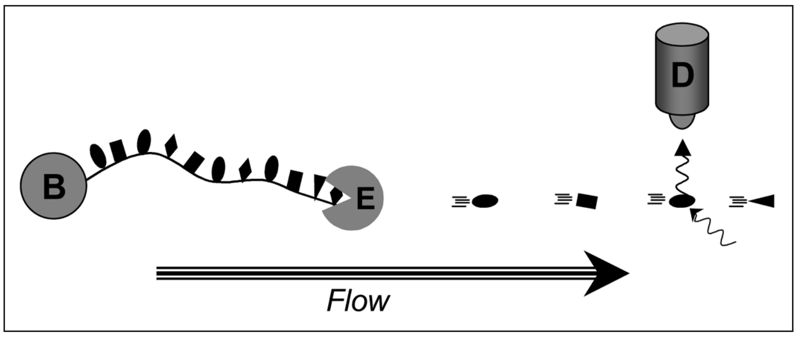

Figure 6: Sequencing by degradation. A DNA molecule consisting of fluorescently modified bases is tethered to a bead (B) held in a flow stream, where consecutive bases are cleaved from one end by a processive exonuclease (E) and carried past a detector (D) which detects and identifies them by their fluorescence

Background fluorescence, complete labelling and efficient nucleotide detection are all obstacles

fluorescently labelled.[93] One of the confounding factors in detecting single nucleotide molecules is the presence of contaminants in almost all solutions, whose intrinsic fluorescence can overwhelm the signal one is trying to detect. This problem has been partly resolved by sophisticated purification and photobleaching methods,[94] but remains an obstacle.

Although considerable progress has been made towards addressing each of the problems with this approach, the successful sequencing of even a short 'model' template remains to be demonstrated. Ultimately, the difficulty in preparing and handling very long, fully labelled DNA templates is likely to be the limiting factor, and may restrict this method to the sequencing of relatively short templates; however, many of the elements of this system may be of great value in other, related approaches.

CONCLUSIONS

Single molecule approaches promise to transform genomics

Single molecule methods, relying either on direct detection and imaging of single molecules or on their *in vitro* amplification, promise to transform genomics over the coming decade.

Already, *in vitro* amplification of single DNA molecules can replace or supplement cell-based cloning in some situations, reducing the risk of biologically induced biases. Cell-free cloning will become more versatile as means are found to amplify larger DNA molecules reliably, preferably by the use of isothermal methods. This will obviate the need for thermocycling, making the growth of 'molecular clones' as simple as that of bacterial ones.

In genome mapping, some single molecule approaches (FISH) are long established. More recent techniques, producing high-resolution, error-free data at high throughput, are already being used in a range of contexts, particularly in providing frameworks for sequencing. But it is single molecule sequencing which poses the greatest challenges and offers the greatest rewards. Within 4–6 years, long reads obtained at high throughput from genomic DNA may enable the rapid sequencing of complex genomes without the need for prior cloning or mapping.

In this paper, the role of microfluidics[95] in single molecule genomics has not been touched on. Every technology — from shipbuilding to watchmaking — needs tools appropriate to its scale. Although the development of microfluidics is being driven mainly by the desire to miniaturise and accelerate conventional genomic techniques, these devices will come into their own as the natural environment in which to work with single molecules.

Finally, what this field of research really needs most is a good name, preferably ending in 'omics'. Telling friends that you are a single molecular geneticist is likely to give entirely the wrong impression.

References

1. Haruna, I. and Spiegelman, S. (1965), 'Specific template requirements of RNA replicases', *Proc. Natl. Acad. Sci. USA*, Vol. 54, pp. 579–587.
2. Blumenthal, T. and Carmichael, G. G. (1979), 'RNA replication: function and structure of Qβ-replicase', *Ann. Rev. Biochem.*, Vol. 48, pp. 525–548.
3. Levisohn, R. and Speigelman, S. (1968), 'The cloning of a self-replicating RNA molecule', *Proc. Natl. Acad. Sci. USA*, Vol. 60, pp. 866–872.
4. Chetverin, A .B., Chetverina, H. V. and Munishkin, A. V. (1991), 'On the nature of spontaneous RNA synthesis by Q beta replicase', *J. Mol. Biol.*, Vol. 222, pp. 3–9.
5. Chetverina, H. V. and Chetverin, A. B. (1993), 'Cloning of RNA molecules in vitro', *Nucleic Acids Res.*, Vol. 21, pp. 2349–2353.
6. Miele, E. A., Mills, D. R. and Kramer, F. R. (1983), 'Autocatalytic replication of a recombinant RNA', *J. Mol. Biol.*, Vol. 171, pp. 281–295.
7. Axelrod, V. D., Brown, E., Priano, C. and Mills, D. R. (1991), 'Coliphage Qβ RNA replication: RNA catalytic for single-strand release', *Virology*, Vol. 184, pp. 595–608.
8. Wu, Y., Zhang, D. Y. and Kramer, F. R. (1992), 'Amplifiable messenger RNA', *Proc. Natl. Acad. Sci. USA*, Vol. 89, pp. 11769–11773.
9. Chetverin, A. B. and Spirin, A. S. (1995), 'Replicable RNA vectors: prospects for cell-

free gene amplification, expression, and cloning', *Prog. Nucleic Acid Res. Mol. Biol.*, Vol. 51, pp. 225–270.

10. Ryabova, L., Volianik, E., Kurnasov, O., Spirin, A. S., Wu, Y. and Kramer, F. R. (1994), 'Coupled replication-translation of amplifiable messenger RNA. A cell-free protein synthesis system that mimics viral infection', *J. Biol. Chem.*, Vol. 269, pp. 1501–1505.

11. Saiki, R. K., Gelfand, D. H., Stoffel, S. *et al.* (1988), 'Primer-directed amplification of DNA with a thermostable DNA polymerase', *Science*, Vol. 239, pp. 487–491.

12. Li, H., Gyllensten, U. B., Cui, X., Saiki, R. K., Erlich, H. A. and Arnheim, N. (1988), 'Amplification and analysis of DNA sequences in single human sperm and diploid cells', *Nature*, Vol. 335, pp. 414–417.

13. Mitra, R. D. and Church, G. M. (1999), '*In situ* localized amplification and contact replication of many individual DNA molecules', *Nucleic Acids Res.*, Vol. 27, e34.

14. Ohuchi, S., Nakano, H. and Yamane, T. (1998), '*In vitro* method for the generation of protein libraries using PCR amplification of single DNA molecule and coupled transcription/translation', *Nucleic Acids Res.*, Vol. 26, pp. 4339–4346.

15. Rungpragayphan, S., Kawarasaki, Y., Imaeda, T., Kohda, K., Nakano, H. and Yamane, T. (2002), 'High-throughput, cloning-independent protein library construction by combining single molecule DNA amplification with *in vitro* expression', *J. Mol. Biol.*, Vol. 318, pp. 395–405.

16. URL: http://www.nwfsc.noaa.gov/protocols/taq-errors.html

17. Andre, P., Kim, A., Khrapko, K. and Thilly, W. G. (1997), 'Fidelity and mutational spectrum of *Pfu* DNA polymerase on a human mitochondrial DNA sequence', *Genome Res.*, Vol. 7, pp. 843–852.

18. Nishioka, M., Mizuguchi, H., Fujiwara, S. *et al.* (2001), 'Long and accurate PCR with a mixture of KOD DNA polymerase and its exonuclease deficient mutant enzyme', *J. Biotech.*, Vol. 88, pp. 141–149.

19. Shizuya, H., Birren, B., Kim, U. J. *et al.* (1992), 'Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in *Escherichia coli* using an F-factor-based vector', *Proc. Natl. Acad. Sci. USA*, Vol. 89, pp. 8794–8797.

20. Burke, D. T., Carle, G. F. and Olson, M. V. (1987), 'Cloning of large segments of exogenous DNA into yeast by means of artificial chromosome vectors', *Science*, Vol. 236, pp. 806–812.

21. Chen, Y. Z., Hayashi, Y., Wu, J. G. *et al.* (2001), 'A BAC-based STS-content map spanning a 35-Mb region of human chromosome 1p35–p36', *Genomics*, Vol. 74, pp. 55–70.

22. Hudson, T. J., Stein, L. D., Gerety, S. S. *et al.* (1995), 'An STS-based map of the human genome', *Science*, Vol. 270, pp. 1945–1954.

23. Gregory, S. G., Soderlund, C. A. and Coulson, A. (1997), 'Contig assembly by fingerprinting', in '*Genome Mapping — A Practical Approach*', Dear, P. H., Ed., IRL Press, Oxford, pp. 227–254.

24. Curran, J. L. (1997), 'Human linkage mapping' in '*Genome Mapping — A Practical Approach*', Dear, P. H., Ed., IRL Press, Oxford. pp. 1–25.

25. Goss, S. J. and Harris, H. (1975), 'New method for mapping genes in human chromosomes', *Nature*, Vol. 255, pp. 680–684.

26. Walter, M. A., Spillett, D. J., Thomas, P., Weissenbach, J. and Goodfellow, P. N. (1994), 'A method for constructing radiation hybrid maps of whole genomes', *Nat. Genet.*, Vol. 7, pp. 22–28.

27. Olivier, M., Aggarwal, A., Allen, J. *et al.* (2001), 'A high-resolution radiation hybrid map of the human genome draft sequence', *Science*, Vol. 291, pp. 1298–1302.

28. Matise, T. C., Porter, C. J., Buyske, S., Cuttichia, J., Sulman, E. P. and White, P. S. (2002), 'Systematic evaluation of map quality: human chromosome 22', *Am. J. Hum. Genet.*, Vol. 70, pp. 1398–1410.

29. Tjio, J. H. and Levan, A. (1956), 'The chromosome number of man', *Am. J. Obstet. Gynecol.*, Vol. 130, pp. 723–724.

30. Lejeune, J., Gautier, M. and Turpin, R. (1959), 'Etude des chromosomes somatiques de neuf enfants mongoliens', *Compt. Rend.*, Vol. 248, pp. 1721–1722.

31. Sturtevant, A. H. (1913), 'The linear arrangement of six sex-linked factors in *Drosophila*, as shown by their mode of association', *J. Exp. Zool.*, Vol. 14, pp. 43–59.

32. Leversha, M. A. (1997), 'Fluorescence *in situ* hybridisation', in '*Genome Mapping — A Practical Approach*', Dear, P. H., Ed., IRL Press, Oxford, pp. 199–225.

33. Yung, J.-F. (1996), 'New FISH probes: the end in sight', *Nat. Genet.*, Vol. 14, pp. 10–12.

34. Trask, B. J. (1991), 'Fluorescence in situ hybridization: applications in cytogenetics and gene-mapping', *Trends Genet.*, Vol. 7, pp. 149–154.

35. Trask, B. J., Pinkel, D. and ven den Engh, G. (1989), 'The proximity of DNA-sequences in interphase cell-nuclei is correlated to genomic distance and permits ordering of cosmids spanning 250 kilobase pairs', *Genomics*, Vol. 5, pp. 710–717.

36. Fidlerova, H., Senger, G., Kost, M., Sanseau,

P. and Sheer, D. (1994), 'Two simple procedures for releasing chromatin from routinely fixed cells for fluorescence *in-situ* hybridization', *Cytogenet. Cell Genet.*, Vol. 65, pp. 203–205.

37. Parra, I. and Windle, B. (1993), 'High-resolution visual mapping of stretched DNA by fluorescent hybridization', *Nat. Genet.*, Vol. 5, pp. 17–21.
38. Schwartz, D. C., Li, X., Hernandez, L. I., Ramnarain, S. P., Huff, E. J. and Wang, Y. K. (1993), 'Ordered restriction maps of *Saccharomyces cerevisiae* chromosomes constructed by optical mapping', *Science*, Vol. 262, pp. 110–114.
39. Cai, W., Aburatani, H., Stanton, V. P., Housman, D. E., Wang, Y.-K. and Schwartz, D. C. (1995), 'Ordered restriction endonuclease maps of yeast artificial chromosomes created by optical mapping on surfaces', *Proc. Natl. Acad. Sci. USA*, Vol. 92, pp. 5164–5168.
40. Lim, A., Dimalanta, E. T., Potamousis, K. D. *et al.* (2001), 'Shotgun optical maps of the whole *Escherichia coli* O157:H7 genome', *Genome Res.*, Vol. 11, pp. 1584–1593.
41. Jing, J. P., Lai, Z. W., Aston, C. *et al.* (1999), 'Optical mapping of *Plasmodium falciparum* chromosome 2', *Genome Res.*, Vol. 9, pp. 175–181.
42. Schäfer, B., Gemeinhardt, H., Uhl, V. and Greulich, K. O. (2000), 'Single molecule DNA restriction analysis in the light microscope', *Single Mol.*, Vol. 1, pp. 33–40.
43. Dear, P. H. and Cook, P. R. (1993), 'HAPPY mapping — linkage mapping using a physical analog of meiosis', *Nucleic Acids Res.*, Vol. 21, pp. 13–20.
44. Dear, P. H. (1997), 'HAPPY mapping', in '*Genome Mapping — A Practical Approach*', Dear, P. H., Ed., IRL Press, Oxford, pp. 94–123.
45. Dear, P. H., Bankier, A. T. and Piper, M. B. (1998), 'A high-resolution metric HAPPY map of human chromosome 14', *Genomics*, Vol. 48, pp. 232–241.
46. Piper, M. B., Bankier, A. T. and Dear, P. H. (1999), 'A HAPPY map of *Cryptosporidium parvum*', *Genome Res.*, Vol. 8, pp. 1299–1307.
47. Konfortov, B. A., Cohen, H. M., Bankier, A. T. and Dear, P. H. (2000), 'A high-resolution HAPPY map of *Dictyostelium discoideum* chromosome 6', *Genome Res.*, Vol. 10, pp. 1737–1742.
48. Glöckner, G., Eichinger, E., Szafranski K. *et al.* (2002), 'Sequence and analysis of chromosome 2 of *Dictyostelium discoideum*', *Nature*, Vol. 418, pp. 79–85.
49. Kostichka, A. J., Marchbanks, M. L., Brummley, RÒ L., Drossman, H. and Smith, L. M. (1992), 'High speed automated DNA sequencing in ultrathin slab gels', *Biotechnology*, Vol. 10, pp. 78–81.
50. Huang, X. C., Quesada, M. A. and Mathies, R. A. (1992), 'DNA sequencing using capillary array electrophoresis', *Anal. Chem.*, Vol. 64, pp. 2149–2154.
51. Meldrum, D. (2000), 'Automation for genomics, part two: sequencers, microarrays, and future trends', *Genome Res.*, Vol. 10, pp. 1288–1303.
52. Paegel, B. M., Emrich, C. A., Weyemayer, G. J., Scherer, J. R. and Mathies, R. A. (2002), 'High throughput DNA sequencing with a microfabricated 96-lane capillary array electrophoresis bioprocessor', *Proc. Natl. Acad. Sci. USA*, Vol. 99, pp. 574–579.
53. Koutny, L., Schmalzing, D., Salas-Solano, O. *et al.* (2000), 'Eight hundred base sequencing in a microfabricated electrophoretic device', *Anal. Chem.*, Vol. 72, pp. 3388–3391.
54. Khrapko, K. R., Lysov, Y. P., Khorlyn, A. A., Shick, V. V., Florentiev, V. L. and Mirzabekov, A. D. (1989), 'An oligonucleotide hybridisation approach to DNA sequencing', *FEBS Lett.*, Vol. 256, pp. 118–122.
55. Southern, E. M. (1996), 'DNA chips: analysing sequence by hybridization to oligonucleotides on a large scale', *Trends Genet.*, Vol. 12, pp. 110–115.
56. Drmanac, S., Kita, D., Labat, I. *et al.* (1998), 'Accurate sequencing by hybridisation for DNA diagnostics and individual genomics', *Nat. Biotechnol.*, Vol. 16, pp. 54–58.
57. Sanger, F., Nicklen, S. and Coulson, A. R. (1977), 'DNA sequencing with chain-terminating inhibitors', *Proc. Natl. Acad. Sci. USA*, Vol. 74, pp. 5463–5467.
58. Ronaghi, M., Nygren, M., Lundeberg, J. and Nyren, P. (1999), 'Analyses of secondary structures in DNA by pyrosequencing', *Anal. Biochem.*, Vol. 267, pp. 65–71.
59. Fakhrai-Rad, H., Pourmand, N. and Ronaghi, M. (2002), 'Pyrosequencing (TM): an accurate detection platform for single nucleotide polymorphisms', *Hum. Mutation*, Vol. 19, pp. 479–485.
60. URL: http://www.pyrosequencing.com
61. Reinartz, J., Bryuns, E., Lin, J.-Z. *et al.* (2002), 'Massively parallel signature sequencing (MPSS) as a tool for in-depth quantitative gene expression profiling in all organisms', *Brief. Func. Genomics Proteomics*, Vol. 1, pp. 95–104.
62. URL: http://www.lynxgen.com
63. Smith, L. M. (1996), 'Sequence from spectrometry: a realistic prospect?', *Nat. Biotechnol.*, Vol. 14, pp. 1084–1087.
64. Oberacher, H., Wellenzohn, B. and Huber, C.

G. (2002), 'Comparative sequencing of nucleic acids by liquid chromatography-tandem mass spectrometry', *Anal. Chem.*, Vol. 74, pp. 211–218.

65. URL: http://www.cbs.dtu.dk/databases/DOGS/

66. Feynman, R. P. (2001), '*The Pleasure of Finding Things Out — The Best Short Works of Richard P. Feynman*', Robbins, J., Ed., Penguin Books, London, p. 125.

67. Poggi, M. A., Bottomley, L. A. and Lillehei, P. T. (2002), 'Scanning probe microscopy', *Anal. Chem.*, Vol. 74, pp. 2851–2862.

68. Mou, J. X., Czajkowsky, D. M., Zhang, Y. Y. and Shao, Z. F. (1995), 'High-resolution atomic force microscopy of DNA: the pitch of the double helix', *FEBS Lett.*, Vol. 371, pp. 279–282.

69. Yoshimura, S. H., Ohniwa, R. L., Sato, M. H. *et al.* (2000), 'DNA phase transition promoted by replication initiator', *Biochemistry*, Vol. 39, pp. 9139–9145.

70. Tanigawa, M., Gotoh, M., Machida, M., Okada, T. and Oishi, M. (2000), 'Detection and mapping of mismatched base pairs in DNA molecules by atomic force microscopy', *Nucleic Acids Res.*, Vol. 28, e38.

71. Woolley, A. T., Guillemette, C., Cheung, C. L., Housman, D. E. and Lieber, C. M. (2000), 'Direct haplotyping of kilobase-size DNA using carbon nanotube probes', *Nat. Biotechnol.*, Vol. 18, pp. 760–763.

72. Rief, M., Clausen-Schaumann, H. and Gaub, H. E. (1999), 'Sequence-dependent mechanics of single molecules', *Nat. Struct. Biol.*, Vol. 6, pp. 346–349.

73. Essevaz-Roulet, B., Bockelmann, U. and Heslot, F. (1997), 'Mechanical separation of the complementary strands of DNA', *Proc. Natl. Acad. Sci. USA*, Vol. 94, pp. 11935–11940.

74. Bockelmann, U., Thomen, P., Essevaz-Roulet, B., Viasnoff, V. and Heslot, F. (2002), 'Unzipping DNA with optical tweezers: high sequence sensitivity and force flips', *Biophys. J.*, Vol. 82, pp. 1537–1553.

75. Deamer, D. W. and Akeson, M. (2000), 'Nanopores and nucleic acids: prospects for ultrarapid DNA sequencing', *Trends Biotech.*, Vol. 18, pp. 147–151.

76. Song, L., Hobaugh, M. R., Shustak, C., Cheley, S., Bayley, H. and Gouaux, J. E. (1996), 'Structure of staphylococcal alpha-hemolysin, a heptameric transmembrane pore', *Science*, Vol. 274, pp. 1859–1866.

77. Kasianowicz, J. J., Brandin, E., Branton, D. and Deamer, D. W. (1996), 'Characterization of individual polynucleotide molecules using a membrane channel', *Proc. Natl. Acad. Sci. USA*, Vol. 93, pp. 13770–13773.

78. Akeson, M., Branton, D., Kasianowicz, J. J., Brandin, E. and Deamer, D. W. (1999), 'Microsecond time-scale discrimination among polycytidylic acid, polyadenylic acid, and polyuridylic acid as homopolymers or as segments within single RNA molecules', *Biophys. J.*, Vol. 77, pp. 3227–3233.

79. Vercoutere, W., Winters-Hilt, S., Olsen, H., Deamer, D., Haussler, D. and Akeson, M. (2001), 'Rapid discrimination among individual DNA hairpin molecules at single-nucleotide resolution using an ion channel', *Nat. Biotechnol.*, Vol. 19, pp. 248–252.

80. Howorka, S., Movileanu, L., Braha, O. and Bayley, H. (2001), 'Kinetics of duplex formation for individual DNA strands within a single protein nanopore', *Proc. Natl. Acad. Sci. USA*, Vol. 98, pp. 12996–13001.

81. Howorka, S., Cheley, S. and Bayley, H. (2001), 'Sequence-specific detection of individual DNA strands using engineered nanopores', *Nat. Biotechnol.*, Vol. 19, pp. 636–639.

82. URL: http://www.solexa.com/

83. Braslavsky, I., Kartalov, E., Hebert, B. and Quake, S. R. (2002), 'Single molecule measurements of DNA polymerase activity: a step towards single molecule sequencing', *Biophys. J.* (Abstract), p. 507.

84. Korlach, J., Levene, M., Turner S. W., Craighead, H. G. and Webb, W. W. (2002), 'Single molecule analysis of DNA polymerase activity using zero-mode waveguides', *Biophys. J.* (Abstracts), p. 507.

85. URL: http://www.nbtc.cornell.edu

86. Jett, J. H., Keller, R. A., Martin, J. C. *et al.* (1989), 'High-speed DNA sequencing: an approach based upon fluorescence detection of single molecules', *J. Biomol. Struct. Dynam.*, Vol. 7, pp. 310–309.

87. Sauer, M., Angerer, B., Ankenbauer, W. *et al.* (2001), 'Single molecule DNA sequencing in submicrometer channels: state of the art and future prospects', *J. Biotechnol.*, Vol. 86, pp. 181–201.

88. Augustin, M. A., Ankenbauer, W. and Angerer, B. (2001), 'Progress towards single-molecule sequencing: enzymatic synthesis of nucleotide-specifically labeled DNA', *J. Biotechnol.*, Vol. 86, pp. 289–301.

89. Földes-Papp, Z., Angerere, B., Ankenbauer, W. and Rigler, R. (2001), 'Fluorescent high-density labeling of DNA: error-free substitution for a normal nucleotide', *J. Biotechnol.*, Vol. 86, pp. 237–253.

90. Földes-Papp, Z., Angerer, B., Thyber, P. *et al.* (2001), 'Fluorescently labeled model DNA sequences for exonucleolytic sequencing', *J. Biotechnol.*, Vol. 86, pp. 203–224.

91. Zhu, Z., Chao, J., Yu, H. and Waggoner,

A. S. (1994), 'Directly labeled DNA probes using fluorescent nucleotides with different length linkers', *Nucleic Acids Res*., Vol. 22, pp. 3418–3422.

92. Werner, J. H., Cai, H., Goodwin, P. M. and Keller, R. A. (1999), 'Current status of DNA sequencing by single molecule detection', *Proc. SPIE*, Vol. 3602, pp. 355–366.

93. Dörre, K., Stephan, J., Lapczyna, M., Stuke, M., Dunkel, H. and Eigen, M. (2001), 'Highly efficient single molecule detection in microstructures', *J. Biotechnol*., Vol. 86, pp. 225–236.

94. Stephan, J., Dörre, K., Brakman, S. *et al*. (2001), 'Towards a general procedure for sequencing single DNA molecules', *J. Biotechnol*., Vol. 86, pp. 255–267.

95. Knight, J. (2002), 'Microfluidics: honey, I shrunk the lab', *Nature*, Vol. 418, pp. 474–475.