ADVANCED SEQUENCING TECHNOLOGIES: METHODS AND GOALS

Jay Shendure, Robi D. Mitra‡, Chris Varma* and George M. Church**

Nearly three decades have passed since the invention of electrophoretic methods for DNA sequencing. The exponential growth in the cost-effectiveness of sequencing has been driven by automation and by numerous creative refinements of Sanger sequencing, rather than through the invention of entirely new methods. Various novel sequencing technologies are being developed, each aspiring to reduce costs to the point at which the genomes of individual humans could be sequenced as part of routine health care. Here, we review these technologies, and discuss the potential impact of such a 'personal genome project' on both the research community and on society.

**Harvard Medical School, 77 Avenue Louis Pasteur, Boston, Massachusetts 02115, USA.*
‡Deptartment of Genetics, Washington University School of Medicine, 4566 Scott Avenue, St. Louis, Missouri 63110, USA.
Correspondence to G.M.C. e-mail address can be found on the following web page: http://arep.med.harvard.edu/gmc
doi:10.1038/nrg1325

The resounding success of the Human Genome Project (HGP) is largely the result of early investments in the development of cost-effective sequencing methods. Over the course of a decade, through the parallelization, automation and refinement of established sequencing methods, the HGP motivated a 100-fold reduction in sequencing costs, from US $10 per finished base to 10 finished bases per US $1 (REF. 1; BOX 1). The relevance and utility of high-throughput sequencing and sequencing centres in the wake of the HGP has been a subject of recent debate. Nonetheless, several academic and commercial efforts are developing new ultra-low-cost sequencing (ULCS) technologies that aim to reduce the cost of DNA sequencing by several orders of magnitude[2,3]. Here, we discuss the motivations for ULCS and review a sample of the technologies themselves.

Until recently, the motivations for pursuing ULCS technologies have generally been defined in terms of the needs and goals of the biomedical and bioagricultural research communities. This list is long, diverse and potentially growing (BOX 2). In more recent years, the primary justification for these efforts has shifted to the idea that the technology could become so affordable that sequencing the full genomes of individual patients would be warranted from a health-care perspective[4–7]. 'Full individual genotyping' has great potential to influence health care, through its contributions to clinical diagnostics and prognostics, risk assessment and disease prevention. Here, we use the phrase 'personal genome project' (PGP) to describe this goal. As we contemplate the routine sequencing of individual human genomes, we must consider the economic, social, legal and ethical issues that are raised by this technology. What are the potential health-care benefits? At what cost threshold does the PGP become viable? What risks does the PGP pose with respect to issues such as consent, confidentiality, discrimination and patient psychology? In addition to reviewing technologies, we try to address several aspects of these questions.

Why continue sequencing?

As a community, we have already sequenced tens of billions of bases and are putting the finishing touches on the canonical human genome. Is a new sequencing technology necessary? Is there anything interesting left to sequence?

Comparative genomics. Through comparative genomics, we are learning a great deal about our own molecular programme, as well as those of other organisms[8,9]. At present, there are more than 3×10^{10} bases in international databases[10]; the genomes of more than 180 organisms

Box 1 | **The first human genome**

In 1977, two groups that were familiar with peptide- and RNA-sequencing methods made a technical leap forward by harnessing the amazing power of gel electrophoresis to separate DNA fragments at single-base resolution[78–81]. In the subsequent decade, electrophoretic sequencing was widely adopted and rapidly improved[82] and, in 1985, a small group of scientists set the audacious goal of sequencing the entire human genome by 2005 (REFS 1,83). The proposal was met with considerable scepticism from the wider community[84,85]; at the time, many felt that the cost of DNA sequencing was far too high (approximately US $10 per base) and that the sequencing community was too fragmented to complete such a vast undertaking. In addition, such 'large-scale biology' represented a significant diversion of resources from the traditional question-driven approach that had been so successful in laying the foundations of molecular biology.

Competition between the Human Genome Project (HGP) and a commercial effort (by Celera) spurred both projects to completion several years ahead of the HGP schedule. Two useful drafts of the human genome were published in 2001 (REFS 85,86). Although the costs of the public project — slightly under US $3 billion — include years of 'production' using weaker technologies, the bulk of the sequencing cost was approximately US $300 million. Among the factors that underlie the achievement of the HGP was the rapid pace of technical and organizational innovation. Crucial factors in achieving the exponential efficiency of sequencing throughput were: automation in the form of commercial sequencing machines, process miniaturization, the optimization of biochemistry and algorithms for sequence assembly. Managerial and organizational challenges were successfully met both within individual sequencing centres and in the way the whole HGP effort was coordinated.

Possibly more significant was the emergence of an 'open' culture with respect to technology, data and software[1]. In refreshing contrast to the competition and consequent secrecy that has traditionally characterized many scientific disciplines, the main sequencing centres freely shared technical advances and engaged in near-instantaneous data release (as formalized in the BERMUDA PRINCIPLES). The approach not only broadened support for the HGP, but also undoubtedly expedited its completion. With respect to both technology development and large-scale biology projects, the HGP perhaps provides us with excellent lessons for how the scientific community can proceed in future endeavours.

have been fully sequenced, as well as parts of the genomes of more than 100,000 taxonomic species[11,12]. It is both humbling and amusing to compare these numbers with the full complexity of the sequences on Earth. By our estimate, a global biomass of more than 2×10^{18} g contains a total biopolymer sequence in the order of 10^{38} residues. From the microbial diversity of the Sargasso Sea[13] to each of the ~6 billion nucleotides of ~6 billion humans, it seems clear that we have only sequenced a small fraction of the full set of interesting and useful nucleotides.

Impact on biomedical research. A widely available ULCS technology would improve current biological and biomedical investigations and expedite the development of several new genomic and technological studies (BOX 2). Foremost among these goals might be efforts to determine the genetic basis of susceptibility to both common and rare human diseases. It is occasionally claimed that all we can afford (and therefore all that we want) to understand so-called multifactorial or complex diseases is information on 'COMMON' SINGLE NUCLEOTIDE POLYMORPHISMS (SNPs), or the arrangements of these (haplotypes)[14,15]. However, all diseases are complex to some degree. Improvements in genotyping and phenotyping methods will increase the chances of finding loci that contribute to ever-lower penetrance and variable expressivity. A focus on common alleles will probably be successful for alleles that are maintained in human populations by heterozygote advantage (such as the textbook relationship between sickle-cell anaemia and malaria) but would miss most of the genetic diseases that have been documented so far[16]. Even for diseases that are amenable to a HAPLOTYPE-MAPPING approach, ULCS would allow geneticists to move more quickly from a haplotype that is linked to a phenotype to the causative SNPs. Diseases that are confounded by GENETIC HETEROGENEITY could be investigated by sequencing specific candidate loci, or whole genomes, across populations of affected individuals[17,18]. It is possible that the cost of accurately genotyping tens of thousands of individuals (for example, US $5,000 for 500,000 SNPs[19] and/or 30,000 genes) will make more sense in the context of routine health care than as stand-alone epidemiology. Whether it occurs by using SNPs or personal genomes, this project will require high levels of informed consent and security[20].

Another broad area that ULCS could influence significantly is cancer biology[21,22]. The ability to sequence and compare complete genomes from many normal, neoplastic and malignant cells would allow us to exhaustively catalogue the molecular pathways and checkpoints that are mutated in cancer. Such a comprehensive approach would help us to more fully decipher the combinations of mutations that together give rise to cancer, and would therefore facilitate a deeper understanding of the cellular functions that are perturbed during tumorigenesis.

ULCS also has the potential to facilitate new research models. Mutagenesis in model and non-model organisms would be more powerful if large genomic regions or complete genomes across large panels of mutant pedigrees could be inexpensively sequenced. In studying acquired immunity, sequencing the rearranged B-cell and T-cell receptor loci in a large panel of lymphocytes could become routine, rather than a large undertaking. ULCS would also benefit the emerging fields of SYNTHETIC BIOLOGY and genome engineering, both of which are becoming powerful tools for perturbing or designing

BERMUDA PRINCIPLES
A commitment that was made in Bermuda (February 1996) by an international assortment of genome-research sponsors to the principles of public sharing and the rapid release of human genome sequence information.

'COMMON' SINGLE NUCLEOTIDE POLYMORPHISMS
(SNPs). Those single-nucleotide substitutions that occur with an allelic frequency of more than 1% in a given population.

HAPLOTYPE MAPPING
A technique that involves the use of combinations of 'common' DNA polymorphisms to find blocks of association with phenotypic traits.

GENETIC HETEROGENEITY
Describes situations in which a similar phenotype can result from various genetic defects.

SYNTHETIC BIOLOGY
A discipline that embraces the emerging ability to design, synthesize and evolve new genomes or biomimetic systems.

PHYLOGENETIC FOOTPRINTING AND SHADOWING
The annotation of functional elements in a genome through bioinformatic comparisons to the genomes of one or more related species.

DIRECTED EVOLUTION
The evolution of a protein (or organism) in the laboratory through rounds of mutation and selection for a particular activity or trait.

FLUORESCENT *IN SITU* SEQUENCING
(FISSEQ). A cyclical, polymerase-driven sequencing method in which nucleotides are modified with fluorescent labels that can be chemically removed at each step.

PHARMACOGENETIC
The heritable component of variation among individuals with respect to drug response or adverse reaction.

Box 2 | **Applications of ultra-low-cost sequencing: a partial list**

- **Sequencing of individual human genomes as a component of preventative medicine.**
- **Rapid hypothesis testing for genotype–phenotype associations[14,17,18].**
- ***In vitro* and *in situ* gene-expression profiling at all stages in the development of a multicellular organism[88,89].**
- **Cancer research: for example, determining comprehensive mutation sets for individual clones[90], carrying out loss-of-heterozygosity analysis[91] and profiling tumour sub-types for diagnosis and prognosis[92,93].**
- **Temporal profiling of B- and T-cell receptor diversity, both clinically and for antibody selection in the laboratory.**
- **Identification of known and new pathogens[94]; development of biowarfare sensors[95].**
- **Detailed annotation of the human genome through PHYLOGENETIC FOOTPRINTING AND SHADOWING[96].**
- **Quantification of alternative splice variants in the transcriptomes of higher eukaryotes[56,97].**
- **Definition of epigenetic structures (such as chromatin modifications and methylation patterns)[98].**
- ***In situ* or *ex vivo* discovery of cell-lineage patterns[99,100].**
- **Characterization of microbial strains that have been subjected to extensive DIRECTED EVOLUTION[101,102].**
- **Exploration of microbial diversity towards agricultural, environmental and therapeutic goals[13,103].**
- **Annotation of microbial genomes through the selectional analysis of tagged insertional mutants[104,105].**
- **Use of DNA or RNA oligonucleotides as agents to bind specific protein targets with high affinity and specificity (so-called 'aptamer technology') for diagnostics and therapeutics[106].**
- **DNA computing[23,24] — that is, manipulating DNA libraries to carry out highly parallel computations. Potential solutions to the problem are often encoded in nucleotide sequence, and standard experimental manipulations (such as hybridization) are used to search the space of possible solutions.**

complex biological systems. This would enable the rapid selection or construction of new enzymes, new genetic networks or perhaps even new chromosomes. DNA computing[23,24] (see BOX 2) and the use of DNA as an ultra-compact means of memory storage loom even farther afield. DNA computing uses only standard recombinant techniques for DNA editing, amplification and detection, but, because these techniques operate on strands of DNA in parallel, the result is a highly efficient process of molecular computing. Furthermore, as 1 g of dehydrated DNA contains approximately 10^{21} bits of information, DNA could potentially store data at a density of 11 orders of magnitude higher than present-day DVDs[24].

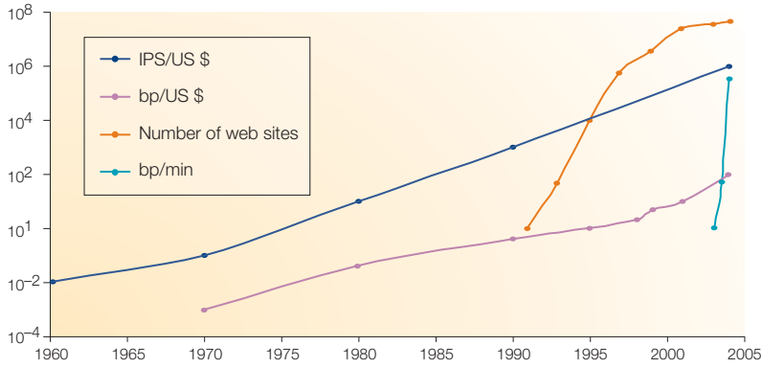

Figure 1 | **Exponential growth in computing and sequencing.** The dark-blue plot indicates the Kurzweil/Moore's Law[108]: it describes the doubling of computer instructions per second per US dollar (IPS/US $) that has been occurring approximately every 18 months since 1900. The magenta plot indicates an exponential growth in the number of base pairs of accurate DNA sequence per unit cost (bp/US $) as a function of time[1]. To some extent, the doubling time for DNA mimics the IPS/US $ curve because it is dependent on it. An even steeper segment occurs in the orange curve; this depicts the number of web sites (doubling time of four months)[109] and shows how quickly a technology can explode when a protocol that can be shared spreads through an existing infrastructure. The turquoise plot is an 'Open Source' case study of 'FLUORESCENT *IN SITU* SEQUENCING' with polonies[40] (see main text for details of this DNA-sequencing technology) in bp/min on simple test templates (doubling time of one month).

The personal genome project. Perhaps the most compelling reason to pursue ULCS technology is the impact that it could have on human health through the sequencing of 'personal genomes' as a component of individualized health care. The current amount of health-care spending for the general population of the United States is approximately US $5,000 per capita per year[25]. Amortized over the 76-year average lifespan for which it is useful, a US $1,000 genome would only have to produce a US $13 benefit per year to break even in terms of cost-effectiveness. Straightforward ways in which full individual genotypes could benefit patient care include clinical diagnostics and prognostics for both common and rare inherited conditions, risk assessment and prevention, and informing patients about any PHARMACOGENETIC contra-indications. Our growing understanding of how specific genotypes and their combinations contribute to the phenome will only increase the value of personal genomes. Even if only rare inherited mutations can be comprehensively surveyed for less than some threshold cost (such as US $5,000), it is probable that each new piece of information that is found in the genome–phenome relationship will make the process more attractive, encouraging the analysis of more genomes and potentially leading to an auto-catalytic shift in the usefulness of personal genomes. The issue now is how this process might get started.

Is the PGP feasible? One reason for the overwhelming success in sequencing the first human genome is that the number of nucleotides that can be sequenced at a given price has increased exponentially over the past 30 years (FIG. 1). This exponential trend is by no means

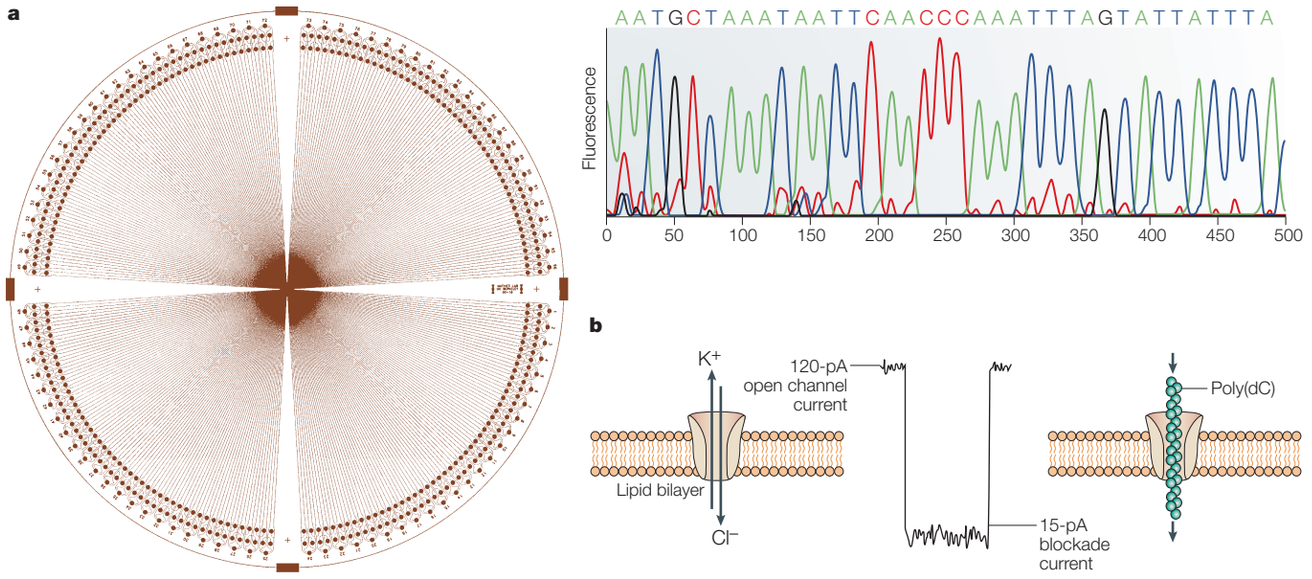

Figure 2 | **Examples of microelectrophoretic sequencing and nanopore sequencing. a** | Microelectrophoretic sequencing. Left: a microfabricated wafer for 384-well capillary electrophoretic sequencing. Reactions are injected at the perimeter and run towards the centre, where a rotary confocal fluorescence scanner carries out the detection. Reproduced with permission from REF. 27 © (2000) *American Chemical Society*. Right: microelectrophoretic sequencing produces raw sequencing traces that are similar to those generated by electrophoretic sequencing[28]. **b** | Nanopore sequencing[57]. Left: single-stranded polynucleotides can only pass single-file through a hemolysin nanopore. Right: the presence of the polynucleotide in the nanopore is detected as a transient blockade of the baseline ionic current. pA, pico-Ampere.

guaranteed, and realizing a PGP in the next five years will probably require a higher commitment to technology development than was available in the pragmatic and production-orientated HGP effort. How might this be achieved? Obviously, we cannot review technologies that are confidential, but several truly innovative approaches have now been made fully or partially public, marking this as an important time to compare and to conceptually integrate these creative strategies. We review five prominant approaches below (see also FIGS 2,3).

Emerging ULCS technologies

Emerging ULCS technologies can be broadly classified into one of five groups: microelectrophoretic methods, 'sequencing by hybridization', cyclic-array sequencing on amplified molecules, cyclic-array sequencing on single molecules and non-cyclical, single-molecule, real-time methods. Most of these technologies are still in the relatively early stages of development, such that it is difficult to gauge when any method will truly be practical and will fulfill expectations. Yet each method has great potential, and several recent technical breakthroughs have contributed to increasing momentum and stimulating community interest in the PGP. To develop a ULCS technology that can deliver low-cost human genomes, it is necessary to take account of the following key parameters: cost per raw base, throughput per instrument, accuracy per raw base and read-length per independent read. With these considerations in mind, BOX 3 outlines the requirements for resequencing a human genome with reasonably high accuracy at a cost of US $1,000.

Microelectrophoretic sequencing. The vast preponderance of DNA sequence has been obtained by using the SANGER-SEQUENCING method, which is based on the electrophoretic separation of deoxyribonucleotide triphosphate (dNTP) fragments with single-base resolution. Using 384-capillary automated sequencing machines, the costs for heavily optimized sequencing centres are currently approaching US $1 per 1,000-bp raw sequencing read and a throughput of ~24 bases per instrument second. Typically, 99.99% accuracy can be achieved with as few as three raw reads covering a given nucleotide. Regions that have proved to be difficult to sequence with conventional protocols can be made accessible through mutagenesis techniques[26]. Several teams, including the Mathies group and researchers at the Whitehead BioMEMS laboratory, are currently investigating whether costs can be further reduced by additional multiplexing and miniaturization[27,28]. By borrowing microfabrication techniques that were developed by the semiconductor industry (FIG. 2a), these groups are working to create single devices that integrate DNA amplification, purification and sequencing[29].

The primary advantage of this approach is that it relies on the same basic principles as electrophoretic sequencing (FIG. 2a), which has already been used to successfully sequence ~10^{11} nucleotides and is therefore well tested. Although the approaches being taken (such as miniaturization and process integration) will certainly yield large cost reductions, achieving 4–5 logs of improvement might require some more radical changes in the underlying engineering of electrophoretic sequencers.

SANGER SEQUENCING
(Chain termination or dideoxy method). A technique that uses an enzymatic procedure to synthesize DNA chains of varying length in four different reactions, stopping the DNA replication at positions that are occupied by one of the four bases, and then determining the resulting fragment lengths to decipher the sequence.

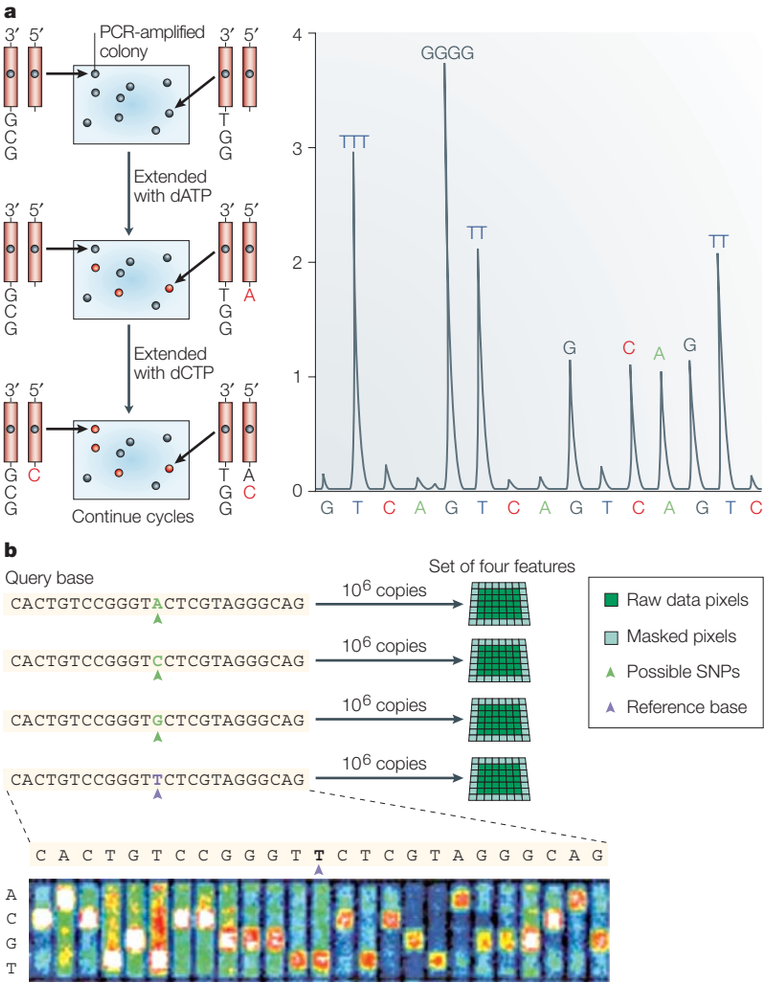

Figure 3 | **Examples of cyclic-array sequencing and sequencing by hybridization.** **a** | Schematic of cyclic-array sequencing-by-synthesis methods (for example, 'fluorescent *in situ* sequencing', Pyrosequencing or single-molecule methods). Left: repeated cycles of polymerase extension with a single nucleotide at each step. The means of detecting incorporation events at individual array features varies from method to method. The sequencing templates shown here have been produced by using the POLONY method[40]. Right: an example of raw data from Pyrosequencing, a cyclic-array method[38]. The identity of nucleotides used at each extension step are listed along the *x*-axis. The *y*-axis depicts the measured signal at each cycle for one sequence; both single and multiple (such as homopolymeric) incorporations can be distinguished from non-incorporation events. The decoded sequence is listed along the top. **b** | Sequencing by hybridization[107]. To resequence a given base, four features are present on the microarray, each identical except for a different nucleotide at the query position (the central base of 25-bp oligonucleotides). Genotyping data at each base are obtained through the differential hybridization of genomic DNA to each set of four features. Part **b** modified with permission from REF. 107 © (2001) CSHL Press.

Nevertheless, given that other ULCS methods are still far from proven, microelectrophoretic sequencing might be a relatively safer option, and might have a higher short-term probability of delivering reasonably low-cost genome resequencing (that is, a 'US $100,000 genome').

Hybridization sequencing. Several efforts are underway to develop sequencing by hybridization (SBH) into a robust and genome-scale sequencing method. The basic principle of SBH is that differential hybridization of oligonucleotide probes can be used to decode a target DNA sequence. One approach is to immobilize the DNA that is to be sequenced on a membrane or glass chip, and then to carry out serial hybridizations with short probe oligonucleotides (for example, 7-bp oligonucleotides). The extent to which specific probes bind the target DNA can be used to infer the unknown sequence. The strategy has been applied to both genome resequencing and *de novo* sequencing[30,31]. Affymetrix and Perlegen have pioneered a different approach to SBH by hybridizing sample DNA to microfabricated arrays of immobilized oligonucleotide probes. The current maximum density of Affymetrix arrays is approximately 1 oligonucleotide 'feature' per 5-μm square; each feature consists of ~100,000 copies of a defined 25-bp oligonucleotide. For each base pair of a reference genome to be resequenced, there are four features on the chip. The middle base pair of these four features is either an A, C, G or T. The sequence that surrounds the variable middle base is identical for all four features and matches the reference sequence (FIG. 3b). By hybridizing labelled sample DNA to the chip and determining which of the four features yields the strongest signal for each base pair in the reference sequence, a DNA sample can be rapidly resequenced. This approach to genome resequencing was first commercialized in the Affymetrix HIV chip in 1995 (REF. 31). Miniaturization, bioinformatics and the availability of a reference human genome sequence allowed Perlegen to greatly extend this approach and develop an oligonucleotide array for resequencing human chromosome 21 (REF. 32). Perlegen has presented unpublished data that extends this approach to the whole genome, but the extent to which the problems discussed below have been addressed is unclear.

SBH technology has a unique set of advantages and challenges. It can be used to obtain an impressive amount of sequence (>10^9 bases) from many distinct chromosomes. Although specific numbers on 'bases per second' are not available, the data-collection method — which involves scanning the fluorescence emitted by labelled target DNA that is hybridized to a wafer array of probe sequences — seems to be compatible with the throughput that is necessary for rapid genome resequencing. For the Affymetrix/Perlegen technology, the effective read-length is set by the length of the query probe (for example, 25 bp, as described in REF. 32). The primary challenges that SBH will face are to design probes or strategies that avoid cross-hybridization of probes to the incorrect targets as a result of repetitive elements or chance similarities. These factors render a substantial fraction of chromosome 21 (>50%) inaccessible[33], and might also contribute to the 3% false-positive SNP-detection rate that was observed in that study. It is also worth noting that SBH still requires sample-preparation steps, as the relevant fraction of the genome must be amplified by PCR before hybridization. In the short term, SBH might have the greatest potential as a technology to query the genotype of a focused set of genomic positions; for example, the ~ten million common SNPs in the human population[34,35].

POLONY
A colony of PCR amplicons that is derived from a single molecule of nucleic acid, amplified *in situ* in an acrylamide gel.

Cyclic-array sequencing on amplified molecules. Cyclic-array methods generally involve multiple cycles of some enzymatic manipulation of an array of spatially separated oligonucleotide features. Each cycle only queries one or a few bases, but thousands to billions of features are processed in parallel. Array features can be ordered or randomly dispersed. Key unifying features of these approaches, including multiplexing in space and time and the avoidance of bacterial clones, emerged as early as 1984 (REF. 36). Early methods in this class led to the first commercially sold genome[37]; however, a dependence on electrophoresis ultimately proved limiting on the speed of data acquisition, and so cyclic sequencing methods that have been developed since then have been non-electrophoretic. In both 'fluorescent *in situ* sequencing' (FISSEQ) and Pyrosequencing, progression through the sequencing reaction is externally controlled by the stepwise (that is, cyclical), polymerase-driven addition of a single type of nucleotide triphosphate to an array of amplified, primed templates. In both cases, repeated cycles of nucleotide extension are used to progressively infer the sequence of individual array features (on the basis of patterns of extension/non-extension over the course of many cycles) (FIG. 3a). Pyrosequencing, which was introduced in 1996, detects extension through the luciferase-based real-time monitoring of pyrophosphate release[38,39]. In FISSEQ, extensions are detected off-line (not in real time) by using the fluorescent groups that are reversibly coupled to deoxynucleotides[40]. Note that both FISSEQ and Pyrosequencing have previously been classified as 'sequencing-by-synthesis' methods. However, as nearly all of the methods reviewed here have crucial synthesis steps, we choose to emphasize cycling as the distinguishing feature of this class.

A third method in this class is based not on cycles of polymerase extension, but instead on cycles of restriction digestion and ligation. In 'massively parallel signature sequencing' (MPSS), array features are sequenced at each cycle by using a TYPE IIS RESTRICTION ENZYME to cleave within a target sequence, leaving a four-base-pair overhang. Sequence-specific ligation of a fluorescent linker is then used to query the identity of the overhang. The achievable 16–20-bp read-lengths (which involve 4–5 cycles) are adequate for many purposes[41].

An additional uniting feature of these methods — one that distinguishes them from several of the single-molecule projects that are discussed below — is that all rely on some method of isolated (that is, clonal) amplification. After amplification, each feature to be sequenced contains thousands to millions of copies of an identical DNA molecule, although features must be spatially distinguishable. The amplification is necessary to achieve sufficient signal for detection. Although the method for clonal amplification is generally independent of the method for cyclic sequencing, all groups seem to have taken different (and creative) routes. In scaling up Pyrosequencing, 454 Corporation used a 'PicoTiter plate' to simultaneously perform hundreds of thousands of picolitre-volume PCR reactions[42]. This was recently applied to the resequencing of the adenovirus genome, but cost and accuracy estimates for this project are not available[43]. For FISSEQ, clonal amplification was achieved by using the polony technology, in which PCR is performed *in situ* in an acrylamide gel[44]. Because the acrylamide restricts the diffusion of the DNA, each single molecule included in the reaction produces a spatially distinct micron-scale colony of DNA (a polony), which can be independently sequenced[45]. For MPSS, each single molecule of DNA in a library is labelled with a unique oligonucleotide tag. After PCR amplification of the library mixture, a proprietary set of paramagnetic 'capture beads' (with each bead bearing an oligonucleotide that is complementary to one of the unique oligonucleotide tags) is used to separate out identical

Box 3 | Is a US $1,000 genome feasible?

To resequence a genome, the sequencing error rate must be significantly lower than the amount of variation that is to be detected[107]. As human chromosomes differ at ~1 in every 1,000 bases, an error rate of 1/100,000 bp is a reasonable goal. If the base accuracy of a RAW READ is ~99.7% (on a par with state-of-the-art instruments), and assuming that errors are random and independent, then ×3 coverage of each base will yield the desired error rate. However, to ensure a minimum ×3 coverage of >95% of a diploid human genome, ~×6.5 coverage is required, or ~40 billion raw bases. In this situation, the cost per base for an accurate US $1,000 genome must approach ~40 million raw bases per US $1 — a 4–5-log improvement over current methods. Although they could potentially approach the cost of a US $2,000 computer, current integrated genomics devices typically cost US $50,000 –500,000. If we assume that the capital/operating costs of our hypothetical instrument are similar to those of conventional electrophoretic sequencers, the bulk of improvements must derive from an increase in the rate of sequence acquisition per device from ~24 bases per second (bp/s) to ~450,000 bp/s. No assembly is required in resequencing a genome; sequencing reads need only be long enough to allow a given read to be matched to a unique location in an assembled reference genome, and then to determine whether and how that read differs from the reference. In a model in which bases are ordered at random, nearly all 20-bp reads would be expected to be unique (4^{20} >>3×10^9). However, as the mammalian genome falls short of being random, only ~73% of 20-bp genomic reads can in fact be assigned to a single unique location. Achieving >95% uniqueness — a modest goal — will require reads of ~60 bp.

Given these assumptions, a resequencing instrument that can deliver a US $1,000 human genome with reasonable coverage and accuracy will need to achieve ~60-bp reads with 99.7% raw-base accuracy, acquiring data at a rate of ~450,000 bp/s. Departures from this situation are almost certain, but will generally involve some trade-off — for example, dropping capital/operating costs by tenfold would enable an instrument with one-tenth of the throughput to achieve the same cost per base.

TYPE IIS RESTRICTION ENZYME
A type of restriction endonuclease that is characterized by an asymmetric recognition site and cleavage at a fixed distance outside the recognition site.

RAW READ
The actual nucleotide sequence that is generated by a sequencing instrument. This contrasts with the finished sequence, which is produced by obtaining the consensus sequence of many overlapping raw reads.

PCR products. The Vogelstein group recently developed a fourth method for achieving clonal amplification, known as BEAM (for beads, emulsion, amplification, magnetic)[46]. In this method, an oil–aqueous emulsion parses a standard PCR reaction into millions of isolated micro-reactors, and magnetic beads are used to capture the clonally amplified products that are generated in individual compartments.

It is worth emphasizing that in the above implementations of cyclic-array sequencing, the methods developed for amplification and sequencing are potentially independent. It is therefore interesting to contemplate possibilities for mixing and matching. For example, it is possible to imagine signature-sequencing polonies, or Pyrosequencing DNA-loaded paramagnetic beads.

The extent to which these methods succeed in realizing ULCS will depend on various factors. Pyrosequencing is close to achieving the required read-lengths, whereas FISSEQ has been shown to achieve reads of only 5–8 bp. Methods that rely on real-time monitoring or manufactured arrays of wells might be difficult to multiplex and miniaturize to the required scale. Crucially, both Pyrosequencing and FISSEQ-based methods must contend with discerning the lengths of homopolymeric sequences — that is, consecutive runs of the same base. Although Pyrosequencing has made significant progress in tackling this challenge through the analysis of the relative amounts of signal that are generated by homopolymers of various lengths (FIG. 3a), the best solution might lie in the development of reversible terminators: these are defined as nucleotides that terminate polymerase extension (for example, through modification of the 3′-hydroxyl group), but that are designed in such a way that the termination properties can be chemically or enzymatically reversed. In addition to circumventing the problem of deciphering homopolymers, reversible terminators would enable simultaneous use all four dNTPs (labelled with different fluorophores). As developing reversible terminators with the necessary properties has proved to be a difficult problem[47,48], recent progress by several groups (described below) is exciting.

Cyclic-array sequencing on single molecules. Each of the methods discussed so far requires either an *in vitro* or *in situ* amplification step, so that the DNA to be sequenced is present at sufficient copy numbers to achieve the required signal. A method for directly sequencing single molecules of DNA would eliminate the need for costly and often problematic procedures, such as cloning and PCR amplification.

Several groups, including Solexa, Genovoxx, Nanofluidics (in collaboration with the Webb group at Cornell University) and Helicos (in collaboration with the Quake group at the California Institute of Technology (Caltech)), are developing cyclic-array methods that are related to those methods discussed above, but that attempt to dispense with the amplification step. Each method relies on the extension of a primed DNA template by a polymerase with fluorescently labelled nucleotides, but they differ in the specifics of their biochemistry and signal detection. In addition, both Solexa and Genovoxx have invested heavily in developing reversibly terminating nucleotides, which would solve the problem (for single-molecule methods as well as amplified cyclic-array methods) of deciphering homopolymeric sequences, by limiting each extension step to a single incorporation. In so far as their research has been revealed at public conferences, Solexa has data on reversible terminators and has shown single-molecule detection with an impressive signal-to-noise ratio. The Genovoxx team has shown the possibility of using standard optics for single-molecule detection and has given details on one class of reversible terminator (C. Hennig, unpublished observations; REF. 48). In the academic sector, the Quake group has recently shown that sequence information can be obtained from single DNA molecules using serial single-base extensions and the clever use of FLUORESCENCE RESONANCE ENERGY TRANSFER (FRET) to improve signal-to-noise ratio[50]. The Webb group has recently shown the real-time detection of nucleotide-incorporation events through a nanofabricated ZERO-MODE WAVEGUIDE. By carrying out the reaction in a zero-mode waveguide, an effective observation volume in the order of 10 zeptolitres (10^{-21} litres) is created, so that in principle, only fluorescent nucleotides in the DNA-polymerase active site are detected[51].

With respect to the ease and reliability of detecting extension events, cyclic-array methods that sequence amplified molecules have an obvious advantage over single-molecule methods. Single-molecule methods, however, have an important advantage in that they avoid a PCR-amplification step, thereby reducing costs and avoiding potential biases (such as sequences that amplify poorly). All methods that are driven by polymerase-based synthesis will probably experience both a low frequency of nucleotide misincorporation and non-incorporation. For amplified-molecule methods, these manifest as eventual signal decay through the 'DEPHASING' of the identical individual templates in a single feature. For single-molecule methods, by contrast, there is no risk of dephasing. A misincorporation event will manifest as a 'dead' template that will not extend further, whereas non-incorporation events will simply appear as a 'pause' in the sequence.

Another advantage of single-molecule methods is that they might require less starting material than other ULCS contenders and conventional sequencing. This feature is relevant to all technologies, and we should take note that methods for amplifying large DNA molecules by MULTIPLE-DISPLACEMENT AMPLIFICATION or WHOLE-GENOME AMPLIFICATION are improving rapidly[52,53]. This will enhance our ability to obtain a complete sequence from single cells even when they are dead or difficult to grow in culture[54,55].

Cyclic-array platforms operate through the spatial separation of single molecules or amplified single molecules. As a consequence of this focus on single molecules, they also allow the determination of combinations of structures that are hard to disentangle in pools of molecules. For example, alternative RNA splicing contributes

FLUORESCENCE RESONANCE ENERGY TRANSFER (FRET). A phenomenon by which excitation is transferred from a donor fluorescent molecule to an acceptor fluorescent molecule; the interaction is distance-dependent and can be used to probe molecular interactions at distances below the limit of optical resolution.

ZERO-MODE WAVEGUIDE
A nanostructure device with physical properties that markedly limit the effective volume of observation.

DEPHASING
The progressive loss of synchronization between templates within features as a consequence of the failure to achieve 100% extension at each extension cycle.

MULTIPLE-DISPLACEMENT AMPLIFICATION
A technique for achieving whole-genome amplification that uses a strand-displacing polymerase to catalyse the isothermal (that is, at a constant temperature) amplification of DNA.

WHOLE-GENOME AMPLIFICATION (WGA). The *in vitro* amplification of a full genome sequence, ideally with even representation of the genome in the amplified product. Techniques for achieving WGA include PCR primed with random or degenerate oligonucleotides, or multiple-displacement amplification.

extensively to protein diversity and regulation, but is poorly assayed by pooled RNAs on microarrays, whereas amplified single molecules allow accurate measures of thousands of alternative spliceforms in RNA molecules, such as those that are transcribed from *CD44* (REF. 56). Similarly, haplotype (or diploid genotype) combinations of SNPs can be determined accurately from DNA molecules (or single cells)[45].

Non-cyclical, single-molecule, real-time methods. A creative single-molecule approach that is unlike all of the above methods is nanopore sequencing, which is being developed by Agilent, and by the Branton and Deamer groups[57–60]. As DNA passes through a 1.5-nm nanopore, different base pairs obstruct the pore to varying degrees, resulting in fluctuations in the electrical conductance of the pore (FIG. 2b). The pore conductance can be measured and used to infer the DNA sequence. The accuracies of base calling range from 60% for single events to 99.9% for 15 events[60]. However, the method has so far been limited to the terminal base pairs of a specific type of hairpin. This method has a great deal of long-term potential for extraordinarily rapid sequencing with little to no sample preparation. However, it is probable that significant pore engineering will be necessary to achieve single-base resolution. Rather than engineering a pore to probe single nucleotides, Visigen and Li-cor are attempting to engineer DNA polymerases or fluorescent nucleotides to provide real-time, base-specific signals while synthesizing DNA at its natural pace (in other words, a non-cyclical sequencing-by-extension system)[61,62].

Implications of sequencing human genomes

Although a thorough consideration of the ethical, legal and social implications of the PGP is available elsewhere[63], we address a few of the issues here.

Clinical advantages and disadvantages. As discussed above, the PGP has the potential to influence patient care in various ways, the most important of which is perhaps by informing diagnostics, prognostics and risk assessment for rare and common diseases that have genetic components. The extent of its usefulness will be a function of the number of genotypes that can be linked to phenotypes. Causative mutations have already been discovered for hundreds of rare conditions[64], and genetic risk factors have been defined for at least ten common diseases[14]. ULCS technology can be expected to accelerate the rate of this discovery. There are also potentially adverse consequences of sequencing a personal genome. Most simply, it might provide more medical information about a patient than the patient wants to know or wants recorded in their medical record. Many patients will not want to know about late-onset diseases, especially if nothing can be done to prevent or ameliorate the condition[63]. Even if laws are passed that prevent genomic information from negatively affecting insurability and employment[65], such laws do not guarantee that an individual's genomic information will never be misused. A debate might therefore rise around the question of whether we should be sequencing whole genomes or restricting data collection or analysis to regions that would be informative to a specific patient's situation[63]. This point seems especially salient with respect to the question of parental rights to sequence the genomes of their children, infants, embryos and fetuses, when the information might or might not be in the subject's best interest[63].

Legal and ethical considerations. With respect to individual subjects, the primary ethical and legal concerns revolve around three main issues[63]: ownership of an individual's DNA and/or its informational content, the purposes for which the information can be used and with whose consent it can be used. In the case of *Moore versus Regents of the University of California*, the court ruled that a patient's informed consent would be required if cells that were removed in the course of their medical treatment were to be used for research. However, the court rejected the idea of property rights to the cells themselves, and that informed consent implies a right to information that is derived from the biological material itself[63]. Fewer than half the states in the United States require informed consent for genetic testing[66], and there are no US federal laws that ban genetic discrimination for medical insurance or in the workplace[65]. More comprehensive protections are probably necessary, but ideally, these should be constructed in such a way that biomedical progress is not impeded. A second category of explicit legal concern is that of patent law. In the United States, Europe and Japan, only portions of DNA that are non-obvious, useful and novel can be patented[67]. ULCS technologies will probably not be able to avoid the resequencing of patented genes. Interesting legal issues arise around the question of patients' rights to have analysed (or to self-analyse) their own DNA sequence versus corporate interests that presumably own the rights to that analysis[63].

Policy and the advancement of science. Beyond vigorously protecting the rights of the individual, we must also consider the welfare of the public with regard to future advancements in biomedicine. Although anonymous data has served the HGP and other biomedical studies well, the approach has limitations. Identity-based genetic information adds significantly to functional genomic studies. As there will be individuals who are willing to make their genome and phenome publicly available, how can comprehensive identifying genetic information be gathered and made available to the research community? There are a few examples of non-anonymous, voluntary public data sets. Craig Venter has published his own genome[68]. Albert Einstein offered his brain for electroencephalography (EEG) and later for neuroanatomy studies[69]. A comprehensive identifying set of COMPUTED TOMOGRAPHY (CT), MAGNETIC RESONANCE (MRI) and serial cryosection images were made from Joseph Jernigan shortly after his execution[70]. Various motivations, ranging from altruism to 'early adopter' technophilia, could arise to encourage individuals to make public their comprehensive identifying data. What

COMPUTED TOMOGRAPHY (CT). An imaging technology that uses computer processing of X-ray images to visualize cross-sectional (transverse) slices of internal structures; the advantage of CT over conventional radiography is the ability to eliminate superimposition.

MAGNETIC RESONANCE [IMAGING] (MRI). A non-invasive technique that is used to generate multi-dimensional proton-density images of internal organs, structures and lesions.

subset of increasingly standardized[71] electronic medical records could such individuals make public? Could these eventually be used to augment expensive epidemiological studies?[72] At present, we have no examples of a publicly available human genome that is coupled to the corresponding phenome[73]. A framework survey and forum for potential volunteers to discuss risks and benefits might be a crucial reality check at this point[74]. Will the response be tiny or will it be as resounding as that following the creation of the Public Library of Science[75], Open Source[76] and Free Software Foundation?[77]

Conclusions

Affordable, personal human genomes as a motivation for developing ULCS technology is a relatively new concept, and one that is only now being viewed as possible in the wake of the HGP. Given where the technologies stand today, and given where they need to be, we should endeavor to be conservative in making projections about when one or more of the ULCS contenders will actually deliver the desired results. It is also important to remember that a significant paradigm shift in sequencing technologies will probably require several years between laboratory proof-of-concept and development of robust commercial systems. Nevertheless, we need to recognize that there have been several recent breakthroughs as well as broadening interest in this field. If the PGP is indeed desirable, then we should start to invest more resources in these technologies straight away[2,3]. ULCS has the potential to catalyse a revolution by bringing genomics to every bedside. Simultaneously, the ready access to genomic information poses potential risks, including breaches of privacy and the misuse of genetic information. In case the PGP does turn out to be just around the corner, we should begin to think clearly about which policy guidelines could best serve the interests of patients, by balancing their right to confidentiality with their need for better medicine.

1. Collins, F. S., Morgan, M. & Patrinos, A. The human genome project: lessons from large-scale biology. *Science* **300**, 286–290 (2003).
A retrospective analysis of the Human Genome Project that was written by the top-level management.
2. National Human Genome Research Institute. *Revolutionary Genome Sequencing Technologies — The $1000 Genome* [online], <http://grants1.nih.gov/grants/guide/rfa-files/RFA-HG-04-003.html> (2004).
3. National Human Genome Research Institute. *Near-term Technology Development for Genome Sequencing* [online], <http://grants.nih.gov/grants/guide/rfa-files/RFA-HG-04-002.html> (2004).
4. Joneitz, E. Personal genomes. *Technol. Rev.* **104**, 30 (2001).
5. Pray, L. A cheap personal genome? *The Scientist* [online], <http://www.biomedcentral.com/news/20021004/04> (2002).
6. Pennisi, E. Gene researchers hunt bargains, fixer-uppers. *Science* **298**, 735–736 (2002).
7. Salisbury, M. W. Fourteen sequencing innovations that could change the way you work. *Genome Technol.* **35**, 40–47 (2003).
8. Carroll, S. B. Genetics and the making of *Homo sapiens*. *Nature* **422**, 840–857 (2003).
9. Ureta-Vidal, A., Ettwiller, L. & Birney, E. Comparative genomics: genome-wide analysis in metazoan eukaryotes. *Nature Rev. Genet.* **4**, 251–62 (2003).
10. National Center for Biotechnology Information. *GenBank Growth* [online], <http://www.ncbi.nlm.nih.gov/Genbank/genbankstats.html> (2003).
11. National Center for Biotechnology Information. *NCBI Taxonomy Browser* [online], <http://www.ncbi.nlm.nih.gov/Taxonomy/txstat.cgi?uncultured=hide&unspecified=hide&m=0> (2004).
12. Integrated Genomics Inc. *Genomes OnLine Database* [online], <http://wit.integratedgenomics.com/GOLD> (2004).
13. Venter, J. C. *et al.* Environmental genome shotgun sequencing of the Sargasso Sea. *Science* 4 Mar 2004 (doi:10.1126/science.1093857).
14. Gibbs, R. A. *et al.* The International HapMap Project. *Nature* **426**, 789–796 (2003).
15. Holtzman N. A. & Marteau T. M. Will genetics revolutionize medicine? *N. Engl. J. Med.* **343**, 141–144 (2000).
16. Vitkup, D., Sander, C. & Church, G. M. The amino-acid mutational spectrum of human genetic disease. *Genome Biol.* **4**, R72 (2003).
17. Farooqi, I. S. *et al.* Clinical spectrum of obesity and mutations in the melanocortin 4 receptor gene. *N. Engl. J. Med.* **348**, 1085–1095 (2003).
18. Smirnova, I. *et al.* Assay of locus-specific genetic load implicates rare Toll-like receptor 4 mutations in meningococcal susceptibility. *Proc. Natl Acad. Sci. USA* **100**, 6075–6080 (2003).
19. Kruglyak L. Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nature Genet.* **22**, 139–144 (1999).
20. Merz J. F., McGee G. E. & Sankar P. 'Iceland Inc.'?: On the ethics of commercial population genomics. *Soc. Sci. Med.* **58**, 1201–1209 (2004).
21. Rajagopalan, H., Nowak, M. A., Vogelstein, B. & Lengauer, C. The significance of unstable chromosomes in colorectal cancer. *Nature Rev. Cancer* **3**, 695–701 (2003).
22. Hanahan, D. & Weinberg, R. A. The hallmarks of cancer. *Cell* **10**, 51–70 (2000).
23. Braich, R. S., Chelyapov, N., Johnson, C., Rothemund, P. W. & Adleman, L. Solution of a 20-variable 3-SAT problem on a DNA computer. *Science* **296**, 499–450 (2002).
24. Reif, J. H. Computing: Successes and challenges. *Science* **296**, 478–479 (2002).
25. Organisation for Economic Cooperation and Development (OECD). *Health Data: Total Expenditure on Health, per capita US$ PPP* [online], <http://www.oecd.org/dataoecd/1/33/2957315. xls> (2003).
26. Keith, J. M. *et al.* Unlocking hidden genomic sequence. *Nucleic Acids Res.* **32**, e35 (2004).
27. Emrich, C. A., Tian, H., Medintz, I. L. & Mathies, R. A. Microfabricated 384-lane capillary array electrophoresis bioanalyzer for ultrahigh-throughput genetic analysis. *Anal. Chem.* **74**, 5076–5083 (2002).
28. Koutny, L. *et al.* Eight hundred-base sequencing in a microfabricated electrophoretic device. *Anal. Chem.* **72**, 3388–3391 (2000).
29. Paegel, B. M., Blazej, R. G. & Mathies, R. A. Microfluidic devices for DNA sequencing: sample preparation and electrophoretic analysis. *Curr. Opin. Biotechnol.* **14**, 42–50 (2003).
A review from the Mathies group on the integration of sample preparation and microelectrophoretic sequencing within a single microfluidic device.
30. Drmanac, S. Accurate sequencing by hybridization for DNA diagnostics and individual genomics. *Nature Biotechnol.* **16**, 54–58 (1998).
31. Drmanac, R. *et al.* DNA sequencing by hybridization with arrays of samples or probes. *Methods Mol. Biol.* **170**, 173–179 (2001).
32. Lipshutz, R. J. *et al.* Using oligonucleotide probe arrays to access genetic diversity. *Biotechniques* **19**, 442–447 (1995).
33. Patil, N., *et al.* Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science* **294**, 1719–1723 (2001).
Perlegen's discovery of SNPs and haplotypes on human chromosome 21 through sequencing by hybridization.
34. Kruglyak, L. & Nickerson, D. A. Variation is the spice of life. *Nature Genet.* **27**, 234–236 (2001).
35. Reich, D. E., Gabriel, S. B. & Altshuler, D. Quality and completeness of SNP databases. *Nature Genet.* **33**, 457–458 (2003).
36. Church, G. M. & Gilbert, W. Genomic sequencing. *Proc. Natl Acad. Sci. USA* **81**, 1991–1995 (1984).
37. Nowak, R. Getting the bugs worked out. *Science* **267**, 172–174 (1995).
38. Ronaghi M. Pyrosequencing sheds light on DNA sequencing. *Genome Res.* **11**, 3–11 (2001).
39. Gharizadeh, B., Nordstrom, T., Ahmadian, A., Ronaghi, M. & Nyren, P. Long-read Pyrosequencing using pure 2′-deoxyadenosine-5′-O′-(1-thiotriphosphate) Sp-isomer. *Analyt. Biochem.* **301**, 82–90 (2002).
Describes the relatively long reads (50–100 bases) obtained through improvements to the Pyrosequencing sequencing-by-synthesis method.
40. Mitra, R. D., Shendure, J., Olejnik, J., Olejnik, E. K. & Church, G. M. Fluorescent *in situ* sequencing on polymerase colonies. *Analyt. Biochem.* **320**, 55–65 (2003).
Introduces the cyclic-array sequencing-by-synthesis technology that is being developed in the Church and Mitra laboratories.
41. Brenner, S. *et al.* Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nature Biotechnol.* **18**, 630–634 (2000).
Describes the 'massively parallel signature sequencing' (MPSS) technology that was developed by Lynx Therapeutics and involves cyclic-array sequencing by serial digestions, ligations and hybridizations.
42. Leamon J. H. *et al.* A massively parallel PicoTiterPlate based platform for discrete picoliter-scale polymerase chain reactions. *Electrophoresis* **24**, 3769–2777 (2003).
43. Sarkis, G. *et al.* Sequence analysis of the pAdEasy-1 recombinant adenoviral construct using the 454 Life Sciences sequencing-by-synthesis method. NCBI AY370911, gi:34014919 (2003).
44. Mitra, R. D. & Church, G. M. *In situ* localized amplification and contact replication of many individual DNA molecules. *Nucleic Acids Res.* **27**, e34 (1999).
45. Mitra, R. D. *et al.* Digital genotyping and haplotyping with polymerase colonies. *Proc. Natl Acad. Sci. USA* **100**, 5926–5931 (2003).
46. Dressman, D., Yan, H., Traverso, G., Kinzler, K. W. & Vogelstein, B. Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations. *Proc. Natl Acad. Sci. USA* **100**, 8817–8822 (2003).
47. Metzker M. L., *et al*. Termination of DNA synthesis by novel 3′-modified-deoxyribonucleoside 5′–triphosphates. *Nucleic Acids Res.* **22**, 4259–4267 (1994).
48. Welch, M. & Burgess, K. Synthesis of fluorescent, photolabile 3′-O-protected nucleoside triphosphates for the base addition sequencing scheme. *Nucleosides Nucleotides* **18**, 197–199 (1999).
49. Hennig, C. AnyBase.nucleotides. *GenoVoxx* [online], <http://www.genovoxx.De> (2004).
50. Braslavsky, I., Hebert, B., Kartalov, E. & Quake, S. R. Sequence information can be obtained from single DNA molecules. *Proc. Natl Acad. Sci. USA* **100**, 3960–3964 (2003).
A description of single-molecule cyclic-array sequencing.
51. Levene, M. J. *et al.* Zero-mode waveguides for single-molecule analysis at high concentrations. *Science* **299**, 682–686 (2003).
Describes the detection of single-nucleotide incorporation events in zeptolitre-scale observation volumes.

52. Dean, F. B. *et al.* Comprehensive human genome amplification using multiple displacement amplification. *Proc. Natl Acad. Sci. USA* **99**, 5261–5266 (2002).
53. Nelson, J. R. *et al.* TempliPhi, phi29 DNA polymerase based rolling circle amplification of templates for DNA sequencing. *Biotechniques* (Suppl.), 44–47 (2002).
54. Sorensen, K. J., Turteltaub, K., Vrankovich, G., Williams, J. & Christian, A. T. Whole-genome amplification of DNA from residual cells left by incidental contact. *Anal. Biochem.* **324**, 312–314 (2004).
55. Rook, M. S., Delach, S. M., Deyneko, G., Worlock, A. & Wolfe, J. L. Whole genome amplification of DNA from laser capture-microdissected tissue for high-throughput single nucleotide polymorphism and short tandem repeat genotyping. *Am. J. Pathol.* **164**, 23–33 (2003).
56. Zhu, J., Shendure, J., Mitra, R. D., & Church, G. M. Single molecule profiling of alternative pre-mRNA splicing. *Science* **301**, 836–838 (2003).
57. Deamer, D. W. & Branton, D. Characterization of nucleic acids by nanopore analysis. *Acc. Chem. Res.* **35**, 817–825 (2002).
58. Li, J., Gershow, M., Stein, D., Brandin, E. & Golovchenko, J. A. DNA molecules and configurations in a solid-state nanopore microscope. *Nature Mater.* **2**, 611–615 (2003).
59. Deamer, D. W. & Akeson, M. Nanopores and nucleic acids: prospects for ultrarapid sequencing. *Trends Biotechnol.* **18**, 147–151 (2000).
A consideration of the successes and remaining challenges of nanopore sequencing.
60. Winters-Hilt S. *et al.* Accurate classification of basepairs on termini of single DNA molecules. *Biophys. J.* **84**, 967–976 (2003).
61. Hardin, S. H. Technologies at VisiGen. *VisiGen Biotechnologies, Inc*. [online], <http://www.visigenbio.com/tech.html> (2004).
62. Williams, J. Heterogenous assay for pyrophosphate. US Patent 6,306,607 (2001).
63. Robertson, J. A. The $1000 genome: ethical and legal issues in whole genome sequencing of individuals. *Am. J. Bioeth.* **3**, W-IF1. (2003).
A well-written overview of the ethical and legal implications of personal genomes.
64. Cooper, D. N. *et al.* Statistics. *Human Gene Mutation Database Cardiff (HGMD)* [online], <http://archive.uwcm.ac.uk/uwcm/mg/docs/hahaha.Html> (2004).
65. Oak Ridge National Laboratory. Genetics Privacy and Legislation. *Human Genome Project Information* [online], <http://www.ornl.gov/TechResources/Human_Genome/elsi/legislat.Html> (2003).
66. Gostin, L. O., Hodge, J. G. & Calvo, C. *Genetics Policy & Law: A Report for Policymakers* (National Council of State Legislators, Washington DC, 2001).
67. Biotechnological Process Patent Act. Pub. L. No. 104–41, 104th Cong., lst Sess. (1 Nov 1995).
68. Venter, J. C. A part of the human genome sequence. *Science* **299**, 1183–1184 (2003).
69. Witelson, S. F., Kigar, D. L. & Harvey, T. The exceptional brain of Albert Einstein. *Lancet* **19**, 2149–2153 (1999).
70. National Library of Medicine. *The Visible Human Project* [online], <http://www.nlm.nih.gov/research/visible/visible_human.Html> (2003).
71. U.S. Government. US Federal Government announcement of first Federal E-Gov Health Information Exchange Standards. *egov Press Releases* [online], <http://www.whitehouse.gov/omb/egov/press/chi_march.Htm> (2003).
72. *Nurses' Health Study at Brigham and Women's Hospital* [online], <http://www.nurseshealthstudy.Org> (2003).
73. Freimer, N. & Sabatti, C. The Human Phenome Project. *Nature Genet.* **3**, 15–21 (2003).
74. Shendure, J. S., Mitra, R. D., Varma, C. & Church, G. M. *Personal Genome Project* [online], <http://arep.med.harvard.edu/PGP> (2004).
75. *Public Library of Science (PLoS)* [online], <http://www.plos.org/cgi-bin/plosSigned.pl> (2004).
76. *SourceForge* [online], <http://sourceforge.net> (2004).
77. *The GNU Prjoect* [online], <http://www.gnu.org> (2004).
78. Sanger, F., Nicklen, S. & Coulson, A. R. DNA sequencing with chain-terminating inhibitors. *Proc. Natl Acad. Sci. USA* **74**, 5463–5467 (1977).
79. Maxam, A. M. & Gilbert, W. A new method for sequencing DNA. *Proc. Natl Acad. Sci. USA* **74**, 560–564 (1977).
80. Gilbert, W. DNA sequencing and gene structure. *Science* **214**, 1305–1312 (1981).
81. Sanger, F. Sequences, sequences, and sequences. *Annu. Rev. Biochem.* **5**, 1–28 (1988).
82. Smith L. M. *et al.* Fluorescence detection in automated DNA sequence analysis. *Nature* **321**, 674–679 (1986).
83. Cook-Deegan, R. M. The Alta summit, December 1984. *Genomics* **5**, 661–663 (1989).
84. Leder P. Can the human genome project be saved from its critics ... and itself? *Cell* **63**,1–3 (1990).
85. Davis B. D. The human genome and other initiatives. *Science* **249**, 342–343 (1990).
86. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
87. Venter, J. C. *et al.* The sequence of the human genome. *Science* **291**, 1304–1351 (2001).
88. Saha, S. *et al.* Using the transcriptome to annotate the genome. *Nature Biotechnol.* **20**, 508–512 (2002).
89. Reymond, A. *et al.* Human chromosome 21 gene expression atlas in the mouse. *Nature* **420**, 582–586 (2002).
90. Hahn, W. C. & Weinberg, R. A. Mechanisms of disease: rules for making human tumor cells. *N. Engl. J. Med.* **34**, 1593–1603 (2002).
91. Paulson, T. G., Galipeau, P. C. & Reid, B. J. Loss of heterozygosity analysis using whole genome amplification, cell sorting, and fluorescence-based PCR. *Genome Res.* **9**, 482–491 (1999).
92. Golub, T. R. *et al.* Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **286**, 531–537 (1999).
93. Ramaswamy, S. *et al.* A molecular signature of metastasis in primary solid tumors. *Nature Genet.* **33**, 49–54 (2003).
94. Weber, G., Shendure, J., Tanenbaum, D. M., Church, G. M., & Meyerson, M. Microbial sequence identification by computational subtraction of the human transcriptome. *Nature Genet.* **30**, 141–142 (2002).
95. Stenger, D. A., Andreadis, J. D., Vora, G. J. & Pancrazio, J. J. Potential applications of DNA microarrays in biodefense-related diagnostics. *Curr. Opin. Biotechnol.* **13**, 208–212 (2002).
96. Boffelli, D. *et al.* Phylogenetic shadowing of primate sequences to find functional regions of the human genome. *Science* **299**, 1391–1394 (2003).
97. Roberts, G. C. & Smith, C. W. Alternative splicing: combinatorial output from the genome. *Curr. Opin. Chem. Biol.* **6**, 375–383 (2002).
98. Robyr, D. *et al.* Microarray deacetylation maps determine genome-wide functions for yeast histone deacetylases. *Cell* **109**, 437–446 (2002).
99. Yatabe, Y., Tavare, S. & Shibata, D. Investigating stem cells in human colon by using methylation patterns. *Proc. Natl Acad. Sci. USA* **9**, 10839–10844 (2001).
100. Dymecki, S. M., Rodriguez, C. I. & Awatramani, R. B. Switching on lineage tracers using site-specific recombination. *Methods Mol. Biol.* **18**, 309–334 (2002).
101. Lenski, R. E., Winkworth, C. L. & Riley, M. A. Rates of DNA sequence evolution in experimental populations of *Escherichia coli* during 20,000 generations. *J. Mol. Evol.* **56**, 498–508 (2003).
102. Cooper, T. F., Rozen, D. E. & Lenski, R. E. Parallel changes in gene expression after 20,000 generations of evolution in *Escherichia coli*. *Proc. Natl Acad. Sci. USA* **100**, 1072–1077 (2003).
103. Gillespie, D. E. *et al.* Isolation of antibiotics turbomycin a and B from a metagenomic library of soil microbial DNA. *Appl. Environ. Microbiol.* **68**, 4301–4306 (2002).
104. Badarinarayana, V. *et al.* Selection analyses of insertional mutants using subgenic-resolution arrays. *Nature Biotechnol.* **1**, 1060–1065 (2001).
105. Sassetti C. M., Boyd D. H. & Rubin E. J. Genes required for mycobacterial growth defined by high density mutagenesis. *Mol. Microbiol.* **48**, 77–84 (2003).
106. Cerchia, L., Hamm, J., Libri, D., Tavitian, B. & de Franciscis, V. Nucleic acid aptamers in cancer medicine. *FEBS Lett.* **528**, 12–16 (2002).
107. Cutler, D. J. *et al.* High-throughput variation detection and genotyping using microarrays. *Genome Res.* **11**, 1913–1925 (2001).
108. Kurzweil, R. The 21st century: a confluence of accelerating revolutions. *KurzweilAI.net* [online], <http://www.kurzweilai.net/meme/frame.html?main=/articles/art0184.html> (2001).
109. Zakon, R. F. *Hobbes' Internet Timeline* [online], <http://www.zakon. org/robert/internet/timeline> (2004).

Acknowledgements

The authors thank members of the polony community and C. Hennig for sharing unpublished work, T. Wu and G. Porreca for helpful discussions, and R. Shendure and K. McKernan for critical reading of the manuscript.

Competing interests statement.

The authors declare competing financial interests: see Web version for details.

Online links

DATABASES

The following terms in this article are linked online to:

LocusLink: http://www.ncbi.nlm.nih.gov/LocusLink
CD44
OMIM: http://www.ncbi.nlm.nih.gov/Omim
malaria | sickle-cell anaemia

FURTHER INFORMATION

International HapMap Project: http://www.hapmap.org
Personal Genome Project: http://arep.med.harvard.edu/PGP
Revolutionary Genome Sequencing Technologies — The $1000 Genome: http://grants1.nih.gov/grants/guide/rfa-files/RFA-HG-04-003.html

Access to this interactive links box is free online.