Published online 12 September 2007

Nucleic Acids Research, 2007, Vol. 35, No. 18 ***6227–6237***
doi:10.1093/nar/gkm688

DNA sequencing: bench to bedside and beyond†

Clyde A. Hutchison III*

J. Craig Venter Institute, 9704 Medical Center Drive, Rockville, MD 20850, USA

Received June 30, 2007; Revised and Accepted August 21, 2007

ABSTRACT

Fifteen years elapsed between the discovery of the double helix (1953) and the first DNA sequencing (1968). Modern DNA sequencing began in 1977, with development of the chemical method of Maxam and Gilbert and the dideoxy method of Sanger, Nicklen and Coulson, and with the first complete DNA sequence (phage ϕX174), which demonstrated that sequence could give profound insights into genetic organization. Incremental improvements allowed sequencing of molecules >200 kb (human cytomegalovirus) leading to an avalanche of data that demanded computational analysis and spawned the field of bioinformatics. The US Human Genome Project spurred sequencing activity. By 1992 the first 'sequencing factory' was established, and others soon followed. The first complete cellular genome sequences, from bacteria, appeared in 1995 and other eubacterial, archaebacterial and eukaryotic genomes were soon sequenced. Competition between the public Human Genome Project and Celera Genomics produced working drafts of the human genome sequence, published in 2001, but refinement and analysis of the human genome sequence will continue for the foreseeable future. New 'massively parallel' sequencing methods are greatly increasing sequencing capacity, but further innovations are needed to achieve the 'thousand dollar genome' that many feel is prerequisite to personalized genomic medicine. These advances will also allow new approaches to a variety of problems in biology, evolution and the environment.

INTRODUCTION

The year 2007 marks the 30th anniversary of the introduction of modern DNA sequencing methods (1,2) and the first complete sequence of a DNA molecule (3,4). These 30 years have seen astounding growth in DNA sequencing capacity and speed. From the first small phage genome, 5386 bases in length, DNA sequencing has advanced to sequence the human genome of ~3 billion bases (5,6). The total amount of sequence in the databases passed the 100 Gb mark in August 2005 (Figure 1). It is remarkable that such progress has been made using methods that are refinements of the basic 'dideoxy' method introduced by Sanger in 1977.

THE SEQUENCE CONCEPT IN BIOLOGY

But the story of sequencing began more than a quarter of a century earlier, when Sanger's studies of insulin first demonstrated the importance of sequence in biological macromolecules. That work showed, for the first time, that proteins are composed of linear polypeptides formed by joining amino acid residues in a defined, but apparently arbitrary order (7,8). Summarizing these studies in his 1959 Nobel address (9), Sanger observed: 'Examination of the sequences of the two chains reveals neither evidence of periodicity of any kind, nor does there seem to be any basic principle which determines the arrangement of the residues. They seem to be put together in an order that is random, but nevertheless unique and most significant, since on it must depend the important physiological action of the hormone.'

Consequently, when the double-helical structure of DNA was proposed soon thereafter (10), it was natural to consider its implications for base sequences in DNA. In their original paper, Watson and Crick pointed out that their structure placed no constraints on the sequence of a DNA molecule. They also observed that it suggested a mechanism for faithful replication of any base sequence. Thus the stage was set for an attack on the coding problem—how is the amino acid sequence of a protein determined by the base sequence of the DNA gene that encodes it? It is interesting that this inferred importance of DNA sequence led to the solution of the coding problem before the experimental determination of any actual DNA sequences, but that is another story.

*To whom correspondence should be addressed. Tel: +1 301 795 7306; Fax: +1 240 268 4004; Email: chutchison@jcvi.org

†A paper by Levy, S., Sutton, G., Ng, P.C., Feuk, L., Halpern, A.L., et al. titled 'The Diploid Genome Sequence of an Individual Human' appeared in PLoS Biology Vol. 5, No. 10, e254 doi:10.1371/journal.pbio.0050254 since the acceptance of this article. That paper reports the diploid genome sequence of J. Craig Venter.

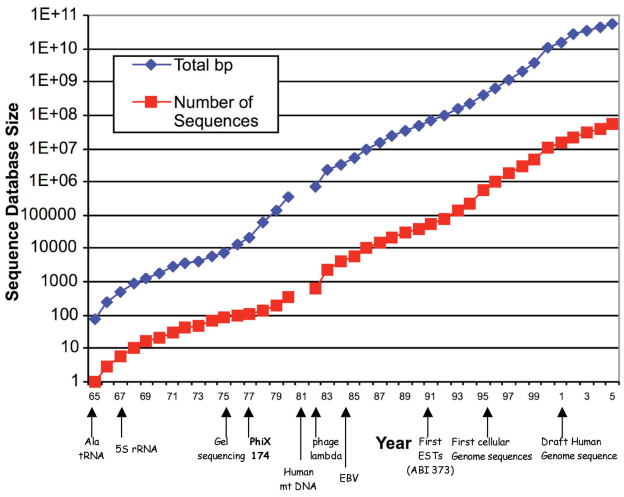

Figure 1. Growth of the nucleotide sequence database. The number of published nucleotide sequences, and the total number of base pairs of sequence are plotted versus the date of deposition or publication. Data since 1981 are re-plotted from http://www.ncbi.nlm.nih.gov/Genbank/genbankstats.html and data for sequences published before 1981 are from Dayhoff, Nucleic Acid Sequence Database, Vol. 1 (38). The dates of landmark sequences and technological advances are indicated.

EARLY DAYS OF DNA SEQUENCING

Fifteen years elapsed between the discovery of the DNA double helix and the first experimental determination of a DNA sequence. This delay was caused by several factors that made the problem intimidating:

(i) The chemical properties of different DNA molecules were so similar that it appeared difficult to separate them.
(ii) The chain length of naturally occurring DNA molecules was much greater than for proteins and made complete sequencing seems unapproachable.
(iii) The 20 amino acid residues found in proteins have widely varying properties that had proven useful in the separation of peptides. The existence of only four bases in DNA therefore seemed to make sequencing a more difficult problem for DNA than for protein.
(iv) No base-specific DNAases were known. Protein sequencing had depended upon proteases that cleave adjacent to certain amino acids.

Some RNA molecules did not share all of these drawbacks with DNA. In particular, transfer RNA molecules were small and individual types could be purified. RNAases with base specificity were known so methods analogous to those used in protein sequencing could be developed. Consequently *Escherichia coli* alanine tRNA was the first nucleic acid molecule to be sequenced by Holley and coworkers in 1965 (11). Unlike amino acid sequences, which were not interpretable until 3D protein structures were determined by X-ray crystallography, models for the structure of tRNAs could be deduced by assuming base pairing analogous to that found in the DNA double helix.

The first DNA molecule purified to homogeneity was the genome of bacteriophage ϕX174, reported by Sinsheimer in 1959 (12). Equilibrium buoyant density centrifugation of the ϕX virion yielded pure preparations from which the DNA could be easily isolated by phenol extraction. ϕX DNA turned out to be a single-stranded circular molecule that was estimated to be ~5000 nt in length.

Purification of other viral DNAs was also accomplished soon, in particular phage lambda DNA (13), a linear molecule with cohesive ends that was the subject of the first successful DNA sequencing. Wu and Kaiser (14,15) measured incorporation of radiolabeled nucleotides by *E. coli* DNA polymerase in reactions that extended the 3′ termini to fill in the complementary cohesive end sequences. They reported a partial sequence in 1968, but completion of the 12 base sequence was not accomplished until 1971 (16). An inroad into DNA sequencing had been made, but the method only applied to short stretches near the ends of lambda and related phage genomes. Wu also provided a means to generalize his approach by introducing the use of oligonucleotide primers in DNA sequencing reactions (17).

The next decisive event, without which DNA sequencing could not have proceeded along the path that it took, was the discovery of type II restriction enzymes by Hamilton Smith and coworkers (18,19). These enzymes recognized and cleaved DNA at specific short nucleotide sequences, typically 4–6 bp in length. It was soon found that enzymes recognizing many different sequences could be identified by screening bacterial strains (20). The restriction enzymes therefore provided a general method for cutting a large DNA molecule into a number of smaller pieces that could be separated by size using gel electrophoresis. These pieces had specific ends that could function as starting points for the sequencing methods that developed over the next few years.

Early work on DNA sequencing that followed sequencing of the phage lambda cohesive ends used methods similar to those used for RNA sequencing. These methods employed base-specific chemical reactions such as depurination, or enzymes with some degree of specificity, for example *E. coli* endonuclease IV, to produce fragments that were typically in the range of 10–20 bp in length. The size of the fragments that could be analyzed was limited by the separation methods that were used. These included 2D chromatography, and also electrophoresis in one dimension followed by chromatography in a second. When such a small fragment was incompletely digested from one end by exonuclease then its sequence could be determined by (i) analysis of the terminal nucleotide of each partial digestion product or (ii) by base-specific shifts in the position of the spot following 2D separation of the products.

Although these early methods were not powerful enough to determine complete gene sequences, several important regulatory signals were sequenced using them. Examples are operator sequences from the *E. coli lac* operon (21) and from phage lambda (22).

GEL-BASED DNA SEQUENCING METHODS

In 1975, Sanger introduced his 'plus and minus' method for DNA sequencing (23,24). This was a critical transition technique leading to the modern generation of methods that have completely dominated sequencing over the past 30 years. The key to this advance was the use of polyacrylamide gels to separate the products of primed synthesis by DNA polymerase in order of increasing chain length. The method analyzed the products of DNA polymerase reactions that extended a primer annealed to a single-stranded DNA template, as Wu and Kaiser had done in sequencing the lambda cohesive ends. DNA synthesis to extend the primer was carried out in two sequential DNA polymerase reactions. The first was carried out under conditions where synthesis was slow and asynchronous, resulting in a population of all possible products extending 1,2,3,... up to a few hundred bases. A ^{32}P labeled nucleotide was incorporated at this step. This product was then divided into eight aliquots and used to prime a second round of DNA polymerase reactions. In these reactions, synthesis was terminated in a sequence-specific manner by supplying only one of the four nucleoside triphosphates ('plus' reactions), or else three of the four ('minus' reactions). The eight reactions were electrophoresed in adjacent lanes of a 12% acrylamide, 8 M urea denaturing gel. Following electrophoresis the gel was placed in contact with X-ray film for a suitable time, typically overnight. When the film was developed, molecules differing by a single nucleotide in length could be resolved as discrete bands on the resulting autoradiograph. This allowed a sequence of ~50 bases to be deduced in a single experiment. The main problem with the method is the difficulty in determining the length of homopolymer runs. Bands corresponding to the beginning and end of such runs are produced, but no bands are produced for positions internal to runs, so run lengths must be estimated from band spacing in the gel. This becomes unreliable for longer runs. Nevertheless, the plus and minus method was used to produce a complete sequence of the ϕX DNA genome (3). In this work, restriction fragments of double-stranded ϕX replicative form DNA were melted and annealed to single-stranded virion DNA to serve as specific primers for the method.

Maxam and Gilbert (2) developed a DNA sequencing method that was similar to the Sanger and Coulson method in using polyacrylamide gels to resolve bands that terminated at each base throughout the target sequence, but very different in the way that products ending in a specific base were generated. Their method started with a double-stranded DNA restriction fragment radiolabeled at one end with ^{32}P. The fragment was then cleaved by base-specific chemical reactions. One reaction cleaves at both purines (the 'A + G' reaction), one preferentially at A ('A > G'), one at pyrimidines ('C + T') and one at cytosines only ('C'). Unlike the plus and minus method, the chemical method produced bands for every sequence position, including those within homopolymer runs. This advantage led to early widespread adoption of the chemical method following its publication in February 1977 (2)

The problems with the plus and minus method were solved when Sanger developed 'the dideoxy method' and published it in December 1977 (1). The underlying concept was to use chain-terminating nucleotide analogs rather than subsets of the four natural dNTPs to cause base-specific termination of primed DNA synthesis. In the original implementation both arabinoside triphosphates and 2′,3′-dideoxy nucleoside triphosphates were tried. These analogs are incorporated in a sequence-specific manner by *E. coli* Pol I, but the enzyme is unable to further extend the growing DNA strand (in the case of the ddNTPs simply because of the lack of a 3′ hydroxyl group). Synthesis was carried out in the presence of all four dNTPs, one of which was α-^{32}P labeled. Four reactions were set up, each doped with a chain-terminating analog of one of the dNTPs, at an appropriate concentration. If the concentration of ddATP, for example, was adjusted so that it was incorporated in place of the normal dATP ~1% of the time then a series of chain-terminated products were produced, each ending with an A. Some molecules in the product ended at each of the A residues in the sequence. When such a product was electrophoresed on a denaturing 12% acrylamide gel, a series of bands representing the positions of all A's in the sequence were displayed. Unlike the plus and minus method, bands were produced for each A within runs of consecutive A residues. When the four dideoxy reactions were run in adjacent lanes it was possible to read sequences of ~100 nt in most cases.

The dideoxy sequencing method as originally described required a single-stranded DNA template. The general applicability of the method was therefore greatly enhanced when Messing and collaborators developed methods for cloning into the single-stranded phage M13 (25–27).

SEQUENCES, SEQUENCES, SEQUENCES

The complete sequence of ϕX determined by the plus and minus method was published in 1977 and then revised slightly in 1978 after resequencing by the dideoxy method. It was a revelation because, to the surprise of many, it turned out to be extremely interesting. Unlike amino acid sequences of proteins, the DNA sequence of the ϕX genome could be interpreted to tell a fascinating story based upon interpretation of the sequence in terms of the genetic code. Analysis of mutations in genes identified by traditional phage genetics, combined with amino acid sequence information for protein components of the ϕX virion, allowed phage genes to be located on the DNA sequence. For the first time translation of a DNA sequence in all possible reading frames identified long open reading frames that could be assigned to genes identified by traditional genetic methods. And, most surprising, it was clear that significant portions of the genome were translated in more than one reading frame to produce two different protein products. These pairs of 'overlapping genes' had not been detected by recombination mapping of the ϕX genome but their existence was indisputable when the sequence was analyzed in light of genetic and protein sequence information (28,29).

The sequence of the simian virus SV40 followed quickly in 1978 (30). Sequencing had begun by determining sequences of RNA copies of parts of the genome, but was rapidly completed after publication of the Maxam–Gilbert method.

With the introduction of the gel-based sequencing methods, the rate of DNA sequencing accelerated (Figure 1). Progress in the methodology was incremental and was driven by the selection of sequencing targets of increasing complexity. In the Sanger group, the 16.5 kb human mitochondrial genome (31) was followed by the 48.5 kb complete phage lambda genome (32). Following Sanger's retirement his protégé Bart Barrell led sequencing of the 172 kb Epstein–Barr virus (33) and then the 237 kb human cytomegalovirus genome (34). For 15 years following ϕX, the sequencing group at the MRC Laboratory of Molecular Biology in Cambridge, UK, had continuously held the record for the longest DNA sequence published.

During this period the useful read length of dideoxy sequencing increased from about 100 up to about 400. This improvement was mainly the result of: (i) the use of very thin sequencing gels and (ii) ^{35}S labeling of the DNA, which gives sharper bands than ^{32}P due to the lower energy of the emitted β particles. Sequencing capacity was also increased by the use of gels with narrow lanes, typically 48 lanes on a 20 by 45 cm gel. Sequencing reactions could be done manually in 96-well plates with handheld repetitive pipetting devices. During this period a single person could run 8 gels on a single day, each with 12 sequence ladders, and obtain some 30 kb of primary sequence data. But it was difficult to do this more than about twice a week.

THE BIRTH OF BIOINFORMATICS

Beginning with ϕX, the management and analysis of sequencing data became a major undertaking. The original ϕX data was in the notebooks of nine different workers each concerned with particular portions of the molecule. Michael Smith, on sabbatical in the Sanger group, had a brother-in-law named Duncan McCallum who was a business computer programmer in Cambridge. He wrote the first programs to help with the compilation and analysis of DNA sequence data (in COBOL) (35). We each transcribed our manually deduced sequences onto paper forms, which then were entered in blocks of 60 on punched cards. The programs then (i) compiled and numbered the complete sequence, (ii) allowed the editing of a previously compiled sequence, (iii) searched the sequence for specific short sequences or families of sequences, for example restriction sites and (iv) translated the sequence in all reading frames. Though invaluable, the programs did not produce output suitable for publication, so the original figure displaying the ϕX sequence with its genes and their translation products annotated (3) was hand typed by Peggy Dowding. Roger Staden helped with computer analysis of the original ϕX sequence and quickly wrote the first suite of programs meant to be interactive and 'designed specifically for use by people with little or no computer experience' (36). These programs developed into the Staden Package, still in use today (37).

With the proliferation of DNA sequence data, came the need for a DNA sequence database. Margaret Dayhoff was the early pioneer in this area. She had previously established a protein sequence database and published the first collection of nucleotide sequence information in 1981 (38). Shortly thereafter, in 1982 GenBank was created by the NIH to provide a 'timely, centralized, accessible repository for genetic sequences' (39).

As the sequence databases grew, methods to compare and align sequences soon became a rate-limiting step in the analysis of sequence data. The development of rapid search programs such as FASTA (40) and BLAST (41) made it practical to identify genes in a new sequence by comparison to all sequences already in the databases.

These are just a few of the early developments in the computer analysis of DNA sequence information. Bioinformatics has developed into a full-blown discipline far beyond the scope of a review such as this. Bioinformatics is central to the interpretation of sequence data and to the generation of testable hypotheses arising from such data.

THE JOHN HENRY SYNDROME

Toward the end of the manual-sequencing era, the first generation of 'automated sequencers' appeared. These machines did not initially automate much of the sequencing process. Gels were still prepared manually, and loaded manually. Only the readout of the sequence data was really automated and in the beginning the base-calling algorithms were quite unsatisfactory. Many steeped in the traditions of manual sequencing were doubtful that automation could compete with dedicated graduate student sequencers. The next few years would show how wrong we were.

AUTOMATED SEQUENCING FACTORIES

In 1986 the laboratory of Leroy Hood at Caltech, in collaboration with Applied Biosystems (ABI), published the first report of automation of DNA sequencing (42). This initial report showed that sequencing data could be collected directly to a computer without autoradiography of the sequencing gel. Although the method could, in principle, have been applied to the chemical sequencing method, the dideoxy method was chosen. A sequencing primer was fluorescently end labeled using four different dyes. A differently labeled primer was used in each of the four dideoxy sequencing reactions. The reactions were combined and electrophoresed in a single polyacrylamide tube gel. DNA was observed by fluorescence as it passed a detector near the bottom of the gel and the four dyes were distinguished by their colors. Fluorescence data was continuously recorded and stored by a computer over the course of a typical 13 h run. The sequence could be deduced from the order in which the four different dyes passed the detector. Work at ABI developed programs to

automatically interpret the data to produce an actual sequence (43).

The ABI 370A DNA sequencer appeared very shortly, and was first used to determine the sequence of a gene by Craig Venter and colleagues at NIH (44). At NIH, Venter set up a sequencing facility with six automated sequencers and two ABI Catalyst robots. In 1992 Venter established The Institute for Genomic Research (TIGR) to expand his sequencing operation and established a facility with 30 ABI 373A automated sequencers and 17 ABI Catalyst 800 robots (45). The organizational differences between this facility and the manual sequencing projects that preceded it were as important as the use of automation. This was a real factory with teams dedicated to different steps in the sequencing process such as template preparation, gel pouring and sequencer operation. Data analysis was integrated into the process so that problems in earlier steps could be detected and corrected as soon as possible. By contrast, in the CMV project, each participating investigator sequenced a region of the genome single-handedly.

An early demonstration of the power of automated sequencing was the development of the expressed sequence tag (EST) approach to gene discovery. In this approach cDNA copies of messenger RNA were cloned at random and subjected to automated sequencing. In the first report from Venter and colleagues in 1991, 337 new human genes were reported, 48 homologous to genes from other organisms (46). This initial study was expanded to an analysis of 83 million nucleotides of cDNA sequence that identified fragments of more than 87 000 human cDNA sequences, more than 77 000 of which were previously unknown (47). This approach was adopted by many genome projects. Today the EST database contains over 43 million ESTs from over 1300 different organisms.

Another early application of the automated sequencer was the worm genome sequencing project (see below) which was underway by 1992 with the beginning elements of a factory atmosphere as well (48).

In 1993 the Sanger Centre, later renamed the Wellcome Trust Sanger Institute, was established by the Wellcome Trust and the Medical Research Council. This is arguably the most important of the sequencing centers established to participate in the worldwide public effort to sequence the human genome. Located in Hinxton, near Cambridge, 'the Sanger' is an outstanding example of a modern sequencing center. The facility has produced $\sim 3.4 \times 10^9$ bases of finished sequence by the 30th anniversary of the dideoxy sequencing method.

CELLULAR GENOMES

Until 1995 the only completely sequenced DNA molecules were viral and organelle genomes. That year Craig Venter's group at TIGR, and their collaborators, reported complete genome sequences of two bacterial species, *Haemophilus influenzae* (49) and *Mycoplasma genitalium* (50). The *H. influenzae* sequence gave the first glimpse of the complete instruction set for a living organism. The *M. genitalium* sequence showed us an approximation to the minimal set of genes required for cellular life.

The methods used to obtain these sequences were as important for subsequent events as the biological insights they revealed. Sequencing of *H. influenzae* introduced the whole genome shotgun (WGS) method for sequencing cellular genomes. In this method, genomic DNA is fragmented randomly and cloned to produce a random library in *E. coli*. Clones are sequenced at random and the results are assembled to produce the complete genome sequence by a computer program that compares all of the sequence reads and aligns matching sequences. Sanger and colleagues used this general strategy to sequence the lambda phage genome (48.5 kb), published in 1982. However, no larger genome was shotgun sequenced until *H. influenzae* (1.83 Mb). In the interim shotgun sequencing was used extensively, but only to sequence mapped subclones of larger sequences. This was the strategy used for the 230 kb human CMV sequence, the largest sequence finished sequencing before the *H. influenzae* genome.

Venter and colleagues introduced critical improvements that made it feasible, for the first time, to shotgun sequence complete cellular genomes. Perhaps most important was adoption of the 'paired ends' strategy (51,52). The automated sequencing procedure used in the *H. influenzae* project used melted double-stranded DNA as template, whereas the HCMV project had used single-stranded M13 vectors. With double-stranded templates it was convenient to sequence each clone from both ends. Because the randomly sheared DNA was carefully sized before cloning, the distance between the reads from the ends of each clone could be estimated. The assembly program used this information to construct 'scaffolds' from the blocks of completely overlapped sequence ('contigs'). When two contigs contained sequences from opposite ends of a single clone, then the two contigs could be linked, although a 'sequence gap' was said to exist between them. Sequence gaps remaining at the end of the shotgun phase of sequencing could be closed by sequencing from a primer for a site internal to a clone bridging the gap. Gaps between scaffolds are 'physical gaps' that contain sequences, which do not occur within any of the sequenced clones. Other measures, such as PCR between the ends of scaffolds using a genomic DNA template, were used to close physical gaps.

Another critical factor in the application of shotgun sequencing to cellular genomes was the TIGR assembler. Previous assembly programs were not designed to handle thousands of sequence reads involved in even the smallest cellular genome projects. However, the TIGR assembler that had been designed to assemble vast amounts of EST data was adequate for the job.

Once these initial sequences were reported the floodgates were open and a steady stream of completed genome sequences has been appearing ever since. It is only possible here to touch on a few of the most significant. Because of the large communities of scientists actively engaged in studies that would benefit from the availability of a genome sequence I have chosen to mention the bacteria *E. coli* and *Bacillus subtilis*, the yeast *Saccharomyces*

cerevisiae, the nematode *Caenorhabditis elegans*, the fruit fly *Drosophila melanogaster* and humans.

Because of its position as the pre-eminent model organism of molecular biology, sequencing of the genome of *E. coli* (4.6 Mb) was proposed by Blattner as early as 1983 (53). Sequencing proceeded as sequencing technology improved, starting with manual methods and finishing in 1997 with automated sequencers (54). Early sequences covering ~1.9 Mb, were deposited starting in 1992, and were obtained from an overlapping set of cosmid clones. The final ~2.5 Mb was obtained by shotgun sequencing of ~250 Kb I-Sce I fragments. This *E. coli* genome sequence, along with several other strains sequenced subsequently has yielded a wealth of information about bacterial evolution and pathogenicity (55,56).

Meanwhile, another model for large-scale genome sequencing projects had emerged; the international consortium. The first genome sequence to be completed by this approach was the yeast *S. cerevisiae* (12.0 Mb) (57), in late 1996. This was the also the first eukaryotic organism to be sequenced. The project involved about 600 scientists in Europe, North America and Japan. The participants included both academic laboratories and large sequencing centers.

The next success of the consortium approach was the genome of the bacterium *B. subtilis* (4.2 Mb) (58), in 1997. The project began in 1990 with the participation of five European laboratories. The project finally became a consortium of 25 laboratories in six European countries coordinated at the Institut Pasteur by Frank Kunst (coordinator) and Antoine Danchin. A consortium of seven Japanese laboratories, coordinated by Naotake Ogasawara and Hiroshi Yoshikawa at the Nara Institute of Science and Technology, Japan, also participated, as well as one Korean and two US laboratories.

The first animal genome sequenced was that of 'the worm' *C. elegans* (97 Mb) (59), in 1998. The authorship of this work was simply 'The *C. elegans* Sequencing Consortium', which was a collaboration between the Washington University Genome Sequencing Center in the United States and the Sanger Centre in UK.

In 1996, ABI introduced the first commercial DNA sequencer that used capillary electrophoresis rather than a slab gel (the ABI Prism 310), and in 1998 the ABI Prism 3700 with 96 capillaries was announced. For the first time DNA sequencing was truly automated. The considerable labor of pouring slab gels was replaced with automated reloading of the capillaries with polymer matrix. Samples for electrophoresis were automatically loaded from 96-well plates rather than manually loaded as the previous generation of sequencers had been. Celera Genomics was found by Applera Corporation (the parent company of ABI) and Craig Venter in May 1998 to exploit these new machines by applying Venter's methods for WGS sequencing to the human genome, in direct competition with the publicly funded Human Genome Project. Celera acquired 300 of the machines, each capable of producing 1.6×10^5 bases of sequence data per day, for a total theoretical capacity of $\sim 5 \times 10^7$ bases of raw sequence data per day.

Celera chose the *D. melanogaster* genome to test the applicability of the WGS approach to a complex eukaryotic genome (60). This involved a scientific collaboration between the scientists at Celera and those of the Berkeley and European *Drosophila* Genome Projects. These projects finished 29 Mb of the 120 Mb of euchromatic portion of the genome. (About one-third of the 180 Mb *Drosophila* genome is centromeric heterochromatin.) Using the WGS approach, data was collected over a 4-month period that provided more than 12× coverage of the euchromatic portion of the genome. The results validated the data produced by the ABI 3700s, the applicability of the WGS approach to eukaryotic genomes, and the assembly methods developed at Celera (61). This was a nearly ideal test case because the WGS data could be analyzed separately and then portions of it could be compared with finished sequence already produced by the *Drosophila* Genome Projects. At the same time the sequence information provided a valuable resource for *Drosophila* genetics. More than 40 scientists at an 'Annotation Jamboree' did initial annotation of the sequence. These scientists, mainly drawn from the *Drosophila* research community, met at Celera for a 2-week period to identify genes, predict functions, and begin a global synthesis of the genome sequence information.

SEQUENCING OF THE HUMAN GENOME

Eventual sequencing of the human genome became an imaginable goal at the outset of the sequencing era 30 years ago. Formal discussions of the idea began in 1985 when Robert Sinsheimer organized a meeting on human genome sequencing at the University of California, Santa Cruz (62). That same year Charles DeLisi and David A. Smith commissioned the first Santa Fe conference, funded by the DOE, to study the feasibility of a Human Genome Initiative. Discussions continued and in 1988 reports recommending a concerted genome research program were issued by committees of the congressional Office of Technology Assessment and the National Research Council. In 1990 the DOE and NIH presented a joint 5-year US Human Genome Project plan to Congress. It was estimated that the project would take 15 years and cost ~3 billion US$.

The US Human Genome Project established goals of mapping, and in some cases sequencing, several model organisms as well as humans. These included *E. coli*, yeast (*S. cerevisiae*), the worm (*C. elegans*), drosophila (*D. melanogaster*) and mouse (laboratory strains of *Mus domesticus*). So, several of the sequences discussed above received funding from the Human Genome Project. The publicly funded effort became an international collaboration between a number of sequencing centers in the United States, Europe and Japan. Each center focused sequencing efforts on particular regions of the genome, necessitating detailed mapping as a first step. In 1994, a detailed genetic map of the human genome was published including 5840 mapped loci with a mean spacing of 0.7 cM (1 centiMorgan = $\sim 10^6$ bp) (63). In 1998 the public project, now in a race with Celera, also adopted the new

ABI Prism 3700 capillary sequencers. In 1999 the Human Genome Project celebrated passing the billion base-pair mark, and the first complete sequence of a human chromosome was reported [chromosome 22 (64)].

Meanwhile at Celera, human genome sequencing was underway using the WGS strategy. Human genome sequencing began in September 1999 and continued until June 2000, when data collection was completed and an initial assembly was achieved. The Celera data provided approximately 5-fold coverage of the genome. An additional 3-fold coverage of unordered and unoriented BAC sequences from the public effort was included in the assembly. The power of the WGS strategy was amply demonstrated.

On 25 June 2000 at the White House, President Clinton with Prime Minister Tony Blair publicly announced draft versions of the human genome sequence from both the publicly funded project and from Celera. In February 2001 the Celera (5) and the public (6) draft human genome sequences were published the same week in *Science* and *Nature*. The race was officially a tie, but it was clear to all that the entry of Celera had speeded the process by several years. Both projects ended up needing the other to make the progress that was made. The Celera assembly benefited from data produced in the public project and the public project quickly adopted some of Celera's methods, in particular the paired-end strategy. Celera's basic methods have now been adopted by all publicly funded genome projects.

Sequencing of the human genome captured public attention in a way that is extremely rare for a scientific topic. Several books for the general public have centered around the 'race' for the human genome sequence (65,66). Leaders of the public and the private projects have even published books describing events from their own personal perspectives (67,68).

BEYOND THE HUMAN GENOME

The major purpose of this review is to commemorate the beginning of the DNA sequencing era with a discussion of the early history of sequencing. But a brief summary of current directions and future horizons is useful to allow an appreciation of the long-term implications of those historical events. It has been said that a full understanding of the human genome sequence may take the better part of the 21st century. Perhaps ironically, the major tool currently available for achieving this understanding is DNA sequencing, on a scale much larger than in the past. Also, we have entered an era where very large-scale DNA sequencing provides a feasible approach to a multitude of problems concerning biology, disease and the environment. This review will conclude with an overview of these developments.

NEXT GENERATION SEQUENCING TECHNOLOGY

In the last few years methods have emerged, which for the first time challenge the supremacy of the dideoxy method. The common feature of these methods is that they are 'massively parallel', meaning that the number of sequence reads from a single experiment is vastly greater than the 96 obtained with modern capillary electrophoresis-based Sanger sequencers. At present this very high throughput is achieved with substantial sacrifices in length and accuracy of the individual reads when compared to Sanger sequencing. Nonetheless, assemblies of such data can be highly accurate because of the high degree of sequence coverage obtainable. The methods are designed for projects that employ the WGS approach. They are most readily applied to resequencing, in which sequence data is aligned with a reference genome sequence in order to look for differences from that reference. A few examples of specific instruments that employ massively parallel strategies are discussed below. Other technologies are under development (69–75), and all of these methods will undoubtedly continue to improve.

The first of the massively parallel methods to become commercially available was developed by 454 Life Sciences (76) and is based on the 'pyrosequencing' technique (77,78). This system allows shotgun sequencing of whole genomes without cloning in *E. coli* or any host cell. First DNA is randomly sheared and ligated to linker sequences that permit individual molecules captured on the surface of a bead to be amplified while isolated within an emulsion droplet (79). A very large collection of such beads is arrayed in the 1.6 million wells of a fiber-optic slide. As with the Sanger method, sequencing is carried out using primed synthesis by DNA polymerase. The array is presented with each of the four dNTPs, sequentially, and the amount of incorporation is monitored by luminometric detection of the pyrophosphate released (hence the name 'pyrosequencing'). A CCD imager coupled to the fiber-optic array collects the data. In sequencing across a homopolymer run, the run length is estimated from the amount of pyrophosphate released, which is proportional to the number of residues incorporated. Errors that result from misjudging the length of homopolymer runs result in single-base insertions and deletions (indels). These constitute the major source of errors in 454 data. The 'plus and minus' method of Sanger and Coulson had this same difficulty. Nonetheless, the second generation 454 Genome Sequencer FLX is reportedly able to produce 100 Mb of sequence with 99.5% accuracy for individual reads averaging read over 250 bases in length.

Another promising technique is the Solexa technology (80,81). A key difference between this method and the 454 is that it uses chain-terminating nucleotides. The fluorescent label on the terminating base can be removed to leave an unblocked 3′ terminus, making chain termination a reversible process. The method reads each base in a homopolymer run in a separate step and therefore does not produce as many indels within such runs as the 454. Because the reversible dye terminator nucleotides are not incorporated efficiently, the read length of the Solexa method is less than for 454. Also more base-substitution errors are observed due to the use of modified polymerase and dye terminator nucleotides. The method sequences clusters of DNA molecules amplified from individual fragments attached randomly on the surface of a flow cell.

Because of the very high densities of clusters that can be analyzed, the machine can reportedly produce 1 billion bases (1 Gb) of 30–40 base sequence reads in a single run.

Applied Biosystems is also developing a massively parallel sequencer, its Supported Oligonucleotide Ligation and Detection system (SOLiD). The technology is based on a hybridization-ligation chemistry (73). The sample preparation aspect of this technology including library preparation, clonal amplification of the target DNA by emulsion PCR on beads is very similar to the 454 processes in principle. However, the size of the beads used for emPCR (1 μm versus 26 μm) and the array format (random versus ordered) are different. These differences afford the SOLiD technology the potential of generating a significantly higher density sequencing array (potentially over a few hundred fold higher), as well as more flexibility in terms of sample input format. The sequence interrogation is done through the repeated cycles of hybridization of a mixture of sequencing primers and fluorescently labeled probes, followed by ligation of the sequencing primers and the probes, then the detection of the fluorescent signals on the probes which encode the bases that are being interrogated. Although it has a short read length of about 25–35, it can generate ~2–3 Gb of sequence per run.

During the development of these new technologies it may be useful to combine sequence data obtained using different techniques. Combination of 454 data with a smaller amount of capillary sequencing data has been shown to be cost-effective in certain situations (82). Similarly, the combination of Solexa data with a smaller amount of 454 data could be effective. The 454 data would provide longer read lengths and a method for obtaining paired end sequences from long DNA fragments to greatly aid assembly (83).

Massively parallel methods of the type described above are likely to dominate high-throughput sequencing applications for the next few years. However, a variety of other approaches are being investigated that may eventually develop into practical methods. Several of these approaches, for example 'nanopore sequencing', have been reviewed (71,84).

GENOMIC MEDICINE

All disease has a genetic basis, whether in genes inherited by the affected individual, environmentally induced genetic changes that produce a cancer, or the genes of a pathogen and their interaction with those of the infected individual. This emerging field of genomic medicine cannot be considered in detail here, but is the subject of several reviews (85–87). Sequencing of the human genome as well as all major pathogens is beginning to have a major impact on the diagnosis, treatment and prevention of diseases. Genome sequences have provided potential targets for drug therapy (88,89) as well as vaccine candidates (90,91).

An era of personalized medicine, informed by information concerning the patient's genotype, has been widely predicted. The Holy Grail in this field has become 'the $1000 genome' sequence. To stimulate work in this area the X Prize Foundation has established the $10 million Archon X Prize for Genomics (92) (http://genomics.xprize.org/).

METAGENOMICS

Only a very small fraction of the microbes found in nature have been grown in pure culture. Consequently we lack a comprehensive view of the genetic diversity to be found on Earth. An approach to this problem has emerged called 'metagenomics' or 'environmental genomics' (93,94). DNA is isolated directly from environmental samples and sequenced, without attempting to culture the organisms from which it comes.

Early studies using manual DNA sequencing methods focused on 16S rRNA genes as a gauge of phylogenetic diversity (95). High-throughput sequencing methods make shotgun sequencing of the whole metagenome from environmental samples informative. A study of the DNA isolated from an acid-mine biofilm yielded nearly complete genome sequences for two organisms and partial sequences for three others from just 76.2 Mb of data (96). Studies of the more complex environment of the oceans have shed new light on the diversity of life on Earth. A total of more than 1.6 Gb of sequence from Sargasso Sea samples yielded 1.2 million previously unknown gene sequences, only about one-third of which were recognizable homologs of previously known genes (97). Before analysis of the Sargasso Sea data the NCBI non-redundant amino acid (nraa) dataset contained some 1.5 million peptides, about 630 000 of which were classified as bacterial. In an expanded study, ocean samples were collected during an expedition from the North Atlantic through the Panama Canal and into the South Pacific. Shotgun sequencing of these samples produced a dataset of 7.7 million sequence reads yielding 6.3 Gb of sequence (98–101). More than 6 million proteins are predicted from this global ocean sampling (GOS) data. These include members of almost all previously known prokaryotic protein families. When these GOS data were clustered with all previously known protein sequences, about 4000 out of a total of 17 000, medium and large clusters contained only GOS sequences (101). The metagenomic approach is being applied to study microbial populations in many environments, for example the human gut (102).

A difficulty with current metagenomic sequence datasets from complex microbial communities is that the vast majority of the data cannot be assembled. This is mainly due to the fact that the cost of acquiring enough sequencing data for the full assembly of all or most of the organisms present in any given sample is still prohibitively high using the state-of-the-art Sanger-based sequencing technology. The development of strategies to apply the new massively parallel sequencing methods may solve this problem if the sequence read length and the accuracy of these technologies can be dramatically improved. In the meantime, another approach is the sequencing of individual cells isolated from the environment. Substantial progress toward this goal has been

achieved (103,104). It is clear that the metagenomic approach is becoming a major new tool for understanding our environment. It will also provide a vast array of new genes for application in the emerging field of synthetic biology.

WHAT NEXT?

The amount of nucleotide sequence in the databases has increased logarithmically by nine orders of magnitude over the 40-year period from 1965 to 2005 (Figure 1). This amounts to an average doubling time of about 16 months. Upon closer inspection, the logarithmic rate of increase is not quite constant over time. Inflection in the curve (Figure 1) appear to correspond to technical innovations, such as the development of the gel-based sequencing techniques 30 years ago, and the introduction of automation. This suggests that we are about to see another inflection in the curve resulting from the next generation of massively parallel sequencers.

In 1977 it was possible to imagine a benchtop machine that could sequence the *E. coli* genome in a few days (although it could only be seen as a black box). The sequencing factory would have been almost inconceivable because it was at odds with the whole culture of biological research. But the sequencing factory did come to pass, and with it came vast amounts of data ideally suited for computational analysis. Taken at face value, the 16-month doubling time for the sequence database is substantially faster than the 24-month doubling in computer power predicted by Moore's law (105) (http://www.intel.com/technology/mooreslaw/). Even if these trends continue it seems unlikely that this difference could begin to limit sequence data collection in the near term. However, some common problems in the analysis of sequence data are already rate-limiting steps in research. A prime example is all-against-all sequence comparison, which requires a computation time that increases with the square of the amount of sequence data. It appears possible that methods for collecting sequence data could soon outstrip our capacity to adequately analyze that data. At present a sizable bioinformatics core is an essential part of a sequencing center. The currently popular vision that an investigator with a single benchtop machine could replace a large sequencing center can only be realized with increases in the productivity of computers and bioinformaticians even more dramatic than that expected for sequencers. It appears that for our individual $1000 genome sequences to be truly useful, fundamental advances in computation and bioinformatics will be essential. The obvious importance of computational analysis of sequence data has led to a greater overall appreciation of the role of theory in biology. A relationship between theory and experiment, not unlike that found in 20th century physics, seems to be taking shape.

ACKNOWLEDGEMENTS

I want to thank Fred Sanger for the privilege of spending a sabbatical year in his laboratory (1975–1976), and for the chance to participate in sequencing the ϕX174 genome. I am grateful for the sabbatical year I spent in Bart Barrell's laboratory (1987–1988) working on the human CMV sequence. I also thank Craig Venter and Ham Smith for our ongoing long-term collaboration, which began in 1995 with the sequencing of the *M. genitalium* genome. These experiences have allowed me to view much of the history of sequencing at close quarters, and may obviously have affected my perspective on a number of, sometimes controversial, issues. I thank Ham Smith for helpful discussions during the writing of this article, concerning matters ranging from the early days of Maxam and Gilbert sequencing to the human genome and beyond. Yu-Hui Rogers has generously shared insights into the new massively parallel sequencing methods. Funding to pay the Open Access publication charges for this article was waived by the editors for this Summary and Survey article.

Conflict of interest statement. None declared.

REFERENCES

1. Sanger,F., Nicklen,S. and Coulson,A.R. (1977) DNA sequencing with chain-terminating inhibitors. *Proc. Natl Acad. Sci. USA*, **74**, 5463–5467.
2. Maxam,A.M. and Gilbert,W. (1977) A new method for sequencing DNA. *Proc. Natl Acad. Sci. USA*, **74**, 560–564.
3. Sanger,F., Air,G.M., Barrell,B.G., Brown,N.L., Coulson,A.R., Fiddes,C.A., Hutchison,C.A., Slocombe,P.M. and Smith,M. (1977) Nucleotide sequence of bacteriophage phi X174 DNA. *Nature*, **265**, 687–695.
4. Sanger,F., Coulson,A.R., Friedmann,T., Air,G.M., Barrell,B.G., Brown,N.L., Fiddes,J.C., Hutchison,C.A.III, Slocombe,P.M. *et al.* (1978) The nucleotide sequence of bacteriophage phiX174. *J. Mol. Biol.*, **125**, 225–246.
5. Venter,J.C., Adams,M.D., Myers,E.W., Li,P.W., Mural,R.J., Sutton,G.G., Smith,H.O., Yandell,M., Evans,C.A. *et al.* (2001) The sequence of the human genome. *Science*, **291**, 1304–1351.
6. Lander,E.S., Linton,L.M., Birren,B., Nusbaum,C., Zody,M.C., Baldwin,J., Devon,K., Dewar,K., Doyle,M. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
7. Sanger,F. (1949) The terminal peptides of insulin. *Biochem. J.*, **45**, 563–574.
8. Sanger,F. and Tuppy,H. (1951) The amino-acid sequence in the phenylalanyl chain of insulin. 2. The investigation of peptides from enzymic hydrolysates. *Biochem. J.*, **49**, 481–490.
9. Sanger,F. (1959) Chemistry of insulin: determination of the structure of insulin opens the way to greater understanding of life processes. *Science*, **129**, 1340–1344.
10. Watson,J.D. and Crick,F.H. (1953) Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid. *Nature*, **171**, 737–738.
11. Holley,R.W., Apgar,J., Everett,G.A., Madison,J.T., Marquisee,M., Merrill,S.H., Penswick,J.R. and Zamir,A. (1965) Structure of a ribonucleic acid. *Science*, **147**, 1462–1465.
12. Sinsheimer,R.L. (1959) A single-stranded DNA from bacteriophage phi X174. *J. Mol. Biol.*, **1**, 43.
13. Kaiser,A.D. and Hogness,D.S. (1960) The transformation of Escherichia coli with deoxyribonucleic acid isolated from bacteriophage lambda-dg. *J. Mol. Biol.*, **2**, 392–415.
14. Kaiser,A.D. and Wu,R. (1968) Structure and function of DNA cohesive ends. *Cold Spring Harb. Symp. Quant. Biol.*, **33**, 729–734.
15. Wu,R. and Kaiser,A.D. (1968) Structure and base sequence in the cohesive ends of bacteriophage lambda DNA. *J. Mol. Biol.*, **35**, 523–537.

16. Wu,R. and Taylor,E. (1971) Nucleotide sequence analysis of DNA. II. Complete nucleotide sequence of the cohesive ends of bacteriophage lambda DNA. *J. Mol. Biol.*, **57**, 491–511.
17. Padmanabhan,R. and Wu,R. (1972) Nucleotide sequence analysis of DNA. IX. Use of oligonucleotides of defined sequence as primers in DNA sequence analysis. *Biochem. Biophys. Res. Commun.*, **48**, 1295–1302.
18. Smith,H.O. and Wilcox,K.W. (1970) A restriction enzyme from *Hemophilus influenzae*. I. Purification and general properties. *J. Mol. Biol.*, **51**, 379–391.
19. Kelly,T.J. Jr and Smith,H.O. (1970) A restriction enzyme from *Hemophilus influenzae*. II. *J. Mol. Biol.*, **51**, 393–409.
20. Middleton,J.H., Edgell,M.H. and Hutchison,C.A.III. (1972) Specific fragments of phi X174 deoxyribonucleic acid produced by a restriction enzyme from *Haemophilus aegyptius*, endonuclease Z. *J. Virol.*, **10**, 42–50.
21. Gilbert,W. and Maxam,A. (1973) The nucleotide sequence of the lac operator. *Proc. Natl Acad. Sci. USA*, **70**, 3581–3584.
22. Maniatis,T., Ptashne,M., Barrell,B.G. and Donelson,J. (1974) Sequence of a repressor-binding site in the DNA of bacteriophage lamda. *Nature*, **250**, 394–397.
23. Sanger,F. and Coulson,A.R. (1975) A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J. Mol. Biol.*, **94**, 441–448.
24. Sanger,F. (1975) The Croonian Lecture, 1975: Nucleotide sequences in DNA. *Proc. R. Soc. Lond. B. Biol. Sci.*, **191**, 317–333.
25. Messing,J., Gronenborn,B., Muller-Hill,B. and Hans Hopschneider,P. (1977) Filamentous coliphage M13 as a cloning vehicle: insertion of a HindII fragment of the lac regulatory region in M13 replicative form in vitro. *Proc. Natl Acad. Sci. USA*, **74**, 3642–3646.
26. Gronenborn,B. and Messing,J. (1978) Methylation of single-stranded DNA in vitro introduces new restriction endonuclease cleavage sites. *Nature*, **272**, 375–377.
27. Heidecker,G., Messing,J. and Gronenborn,B. (1980) A versatile primer for DNA sequencing in the M13mp2 cloning system. *Gene*, **10**, 69–73.
28. Barrell,B.G., Air,G.M. and Hutchison,C.A.III. (1976) Overlapping genes in bacteriophage phiX174. *Nature*, **264**, 34–41.
29. Smith,M., Brown,N.L., Air,G.M., Barrell,B.G., Coulson,A.R., Hutchison,C.A.III. and Sanger,F. (1977) DNA sequence at the C termini of the overlapping genes A and B in bacteriophage phi X174. *Nature*, **265**, 702–705.
30. Fiers,W., Contreras,R., Haegemann,G., Rogiers,R., Van de Voorde,A., Van Heuverswyn,H., Van Herreweghe,J., Volckaert,G. and Ysebaert,M. (1978) Complete nucleotide sequence of SV40 DNA. *Nature*, **273**, 113–120.
31. Anderson,S., Bankier,A.T., Barrell,B.G., de Bruijn,M.H., Coulson,A.R., Drouin,J., Eperon,I.C., Nierlich,D.P., Roe,B.A. *et al.* (1981) Sequence and organization of the human mitochondrial genome. *Nature*, **290**, 457–465.
32. Sanger,F., Coulson,A.R., Hong,G.F., Hill,D.F. and Petersen,G.B. (1982) Nucleotide sequence of bacteriophage lambda DNA. *J. Mol. Biol.*, **162**, 729–773.
33. Baer,R., Bankier,A.T., Biggin,M.D., Deininger,P.L., Farrell,P.J., Gibson,T.J., Hatfull,G., Hudson,G.S., Satchwell,S.C. *et al.* (1984) DNA sequence and expression of the B95-8 Epstein–Barr virus genome. *Nature*, **310**, 207–211.
34. Bankier,A.T., Beck,S., Bohni,R., Brown,C.M., Cerny,R., Chee,M.S., Hutchison,C.A.III, Kouzarides,T., Martignetti,J.A. *et al.* (1991) The DNA sequence of the human cytomegalovirus genome. *DNA Seq.*, **2**, 1–12.
35. McCallum,D. and Smith,M. (1977) Computer processing of DNA sequence data. *J. Mol. Biol.*, **116**, 29–30.
36. Staden,R. (1977) Sequence data handling by computer. *Nucleic Acids Res.*, **4**, 4037–4051.
37. Staden,R., Beal,K.F. and Bonfield,J.K. (2000) The Staden package, 1998. *Methods Mol. Biol.*, **132**, 115–130.
38. Dayhoff,M.O., Schwartz,R.M., Chen,H.R., Hunt,L.T., Barker,W.C. and Orcutt,B.C. (1981) *Nucleic acid sequence database,* Vol. **1**. National Biomedical Research Foundation, Georgetown University Medical Center, Washington D.C. 20001.
39. Bilofsky,H.S., Burks,C., Fickett,J.W., Goad,W.B., Lewitter,F.I., Rindone,W.P., Swindell,C.D. and Tung,C.S. (1986) The GenBank genetic sequence databank. *Nucleic Acids Res.*, **14**, 1–4.
40. Pearson,W.R. and Lipman,D.J. (1988) Improved tools for biological sequence comparison. *Proc. Natl Acad. Sci. USA*, **85**, 2444–2448.
41. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
42. Smith,L.M., Sanders,J.Z., Kaiser,R.J., Hughes,P., Dodd,C., Connell,C.R., Heiner,C., Kent,S.B. and Hood,L.E. (1986) Fluorescence detection in automated DNA sequence analysis. *Nature*, **321**, 674–679.
43. Connell,C., Fung,S., Heiner,C., Bridgham,J., Chakerian,V., Heron,E., Jones,B., Menchen,S., Mprdan,W. *et al.* (1987) Automated DNA-Sequence analysis. *Biotechniques*, **5**, 342–348.
44. Gocayne,J., Robinson,D.A., FitzGerald,M.G., Chung,F.Z., Kerlavage,A.R., Lentes,K.U., Lai,J., Wang,C.D., Fraser,C.M. *et al.* (1987) Primary structure of rat cardiac beta-adrenergic and muscarinic cholinergic receptors obtained by automated DNA sequence analysis: further evidence for a multigene family. *Proc. Natl Acad. Sci. USA*, **84**, 8296–8300.
45. Adams,M.D., Kerlavage,A.R., Kelley,J.M., Gocayne,J.D., Fields,C., Fraser,C.M. and Venter,J.C. (1994) A model for high-throughput automated DNA sequencing and analysis core facilities. *Nature*, **368**, 474–475.
46. Adams,M.D., Kelley,J.M., Gocayne,J.D., Dubnick,M., Polymeropoulos,M.H., Xiao,H., Merril,C.R., Wu,A., Olde,B. *et al.* (1991) Complementary DNA sequencing: expressed sequence tags and human genome project. *Science*, **252**, 1651–1656.
47. Adams,M.D., Kerlavage,A.R., Fleischmann,R.D., Fuldner,R.A., Bult,C.J., Lee,N.H., Kirkness,E.F., Weinstock,K.G., Gocayne,J.D. *et al.* (1995) Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequence. *Nature*, **377**, 3–174.
48. Sulston,J., Du,Z., Thomas,K., Wilson,R., Hillier,L., Staden,R., Halloran,N., Green,P., Thierry-Mieg,J. *et al.* (1992) The *C. elegans* genome sequencing project: a beginning. *Nature*, **356**, 37–41.
49. Fleischmann,R.D., Adams,M.D., White,O., Clayton,R.A., Kirkness,E.F., Kerlavage,A.R., Bult,C.J., Tomb,J.F., Dougherty,B.A. *et al.* (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science*, **269**, 496–512.
50. Fraser,C.M., Gocayne,J.D., White,O., Adams,M.D., Clayton,R.A., Fleischmann,R.D., Bult,C.J., Kerlavage,A.R., Sutton,G. *et al.* (1995) The minimal gene complement of *Mycoplasma genitalium*. *Science*, **270**, 397–403.
51. Edwards,A., Voss,H., Rice,P., Civitello,A., Stegemann,J., Schwager,C., Zimmermann,J., Erfle,H., Caskey,C.T. *et al.* (1990) Automated DNA sequencing of the human HPRT locus. *Genomics*, **6**, 593–608.
52. Roach,J.C., Boysen,C., Wang,K. and Hood,L. (1995) Pairwise end sequencing: a unified approach to genomic mapping and sequencing. *Genomics*, **26**, 345–353.
53. Blattner,F.R. (1983) Biological Frontiers. *Science*, **222**, 2.
54. Blattner,F.R., Plunkett,G.III, Bloch,C.A., Perna,N.T., Burland,V., Riley,M., Collado-Vides,J., Glasner,J.D., Rode,C.K. *et al.* (1997) The complete genome sequence of *Escherichia coli* K-12. *Science*, **277**, 1453–1474.
55. Perna,N.T., Plunkett,G.III, Burland,V., Mau,B., Glasner,J.D., Rose,D.J., Mayhew,G.F., Evans,P.S., Gregor,J. *et al.* (2001) Genome sequence of enterohaemorrhagic *Escherichia coli* O157:H7. *Nature*, **409**, 529–533.
56. Welch,R.A., Burland,V., Plunkett,G.III, Redford,P., Roesch,P., Rasko,D., Buckles,E.L., Liou,S.R., Boutin,A. *et al.* (2002) Extensive mosaic structure revealed by the complete genome sequence of uropathogenic *Escherichia coli*. *Proc. Natl Acad. Sci. USA*, **99**, 17020–17024.
57. Goffeau,A., Barrell,B.G., Bussey,H., Davis,R.W., Dujon,B., Feldmann,H., Galibert,F., Hoheisel,J.D., Jacq,C. *et al.* (1996) Life with 6000 genes. *Science*, **274**, 546, 563–547.
58. Kunst,F., Ogasawara,N., Moszer,I., Albertini,A.M., Alloni,G., Azevedo,V., Bertero,M.G., Bessieres,P., Bolotin,A. *et al.* (1997) The complete genome sequence of the gram-positive bacterium *Bacillus subtilis*. *Nature*, **390**, 249–256.
59. The C. elegans Sequencing Consortium. (1998) Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science*, **282**, 2012–2018.

60. Adams,M.D., Celniker,S.E., Holt,R.A., Evans,C.A., Gocayne,J.D., Amanatides,P.G., Scherer,S.E., Li,P.W., Hoskins,R.A. *et al.* (2000) The genome sequence of *Drosophila melanogaster*. *Science*, **287**, 2185–2195.
61. Myers,E.W., Sutton,G.G., Delcher,A.L., Dew,I.M., Fasulo,D.P., Flanigan,M.J., Kravitz,S.A., Mobarry,C.M., Reinert,K.H. *et al.* (2000) A whole-genome assembly of Drosophila. *Science*, **287**, 2196–2204.
62. Sinsheimer,R.L. (2006) To reveal the genomes. *Am. J. Hum. Genet.*, **79**, 194–196.
63. Murray,J.C., Buetow,K.H., Weber,J.L., Ludwigsen,S., Scherpbier-Heddema,T., Manion,F., Quillen,J., Sheffield,V.C., Sunden,S. *et al.* (1994) A comprehensive human linkage map with centimorgan density. Cooperative Human Linkage Center (CHLC). *Science*, **265**, 2049–2054.
64. Dunham,I., Shimizu,N., Roe,B.A., Chissoe,S., Hunt,A.R., Collins,J.E., Bruskiewich,R., Beare,D.M., Clamp,M. *et al.* (1999) The DNA sequence of human chromosome 22. *Nature*, **402**, 489–495.
65. Wickelgren,I. (2002) edn. *The Gene Masters: How a New Breed of Scientific Entrepreneurs Raced for the Biggest Prize in Biology, 1st edn*. Times Books/Henry Holt and Co., New York.
66. Shreeve,J. (2004) *The Genome War: How Craig Venter Tried to Capture the Code of Life and Save the World*. Alfred A. Knopf, New York.
67. Collins,F.S. (2006) *The Language of God: A Scientist Presents Evidence for Belief*. Free Press, New York.
68. Venter,J.C. (2007, in the press) *A Life Decoded*. Viking Press, New York.
69. Metzker,M.L. (2005) Emerging technologies in DNA sequencing. *Genome Res.*, **15**, 1767–1776.
70. Rogers,Y.H. and Venter,J.C. (2005) Genomics: massively parallel sequencing. *Nature*, **437**, 326–327.
71. Chan,E.Y. (2005) Advances in sequencing technology. *Mutat. Res.*, **573**, 13–40.
72. Braslavsky,I., Hebert,B., Kartalov,E. and Quake,S.R. (2003) Sequence information can be obtained from single DNA molecules. *Proc. Natl Acad. Sci. USA*, **100**, 3960–3964.
73. Shendure,J., Porreca,G.J., Reppas,N.B., Lin,X., McCutcheon,J.P., Rosenbaum,A.M., Wang,M.D., Zhang,K., Mitra,R.D. *et al.* (2005) Accurate multiplex polony sequencing of an evolved bacterial genome. *Science*, **309**, 1728–1732.
74. Blazej,R.G., Kumaresan,P. and Mathies,R.A. (2006) Microfabricated bioprocessor for integrated nanoliter-scale Sanger DNA sequencing. *Proc. Natl Acad. Sci. USA*, **103**, 7240–7245.
75. Rich,A. (1998) The rise of single-molecule DNA biochemistry. *Proc. Natl Acad. Sci. USA*, **95**, 13999–14000.
76. Margulies,M., Egholm,M., Altman,W.E., Attiya,S., Bader,J.S., Bemben,L.A., Berka,J., Braverman,M.S., Chen,Y.J. *et al.* (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, **437**, 376–380.
77. Nyren,P., Pettersson,B. and Uhlen,M. (1993) Solid phase DNA minisequencing by an enzymatic luminometric inorganic pyrophosphate detection assay. *Anal. Biochem.*, **208**, 171–175.
78. Ronaghi,M., Karamohamed,S., Pettersson,B., Uhlen,M. and Nyren,P. (1996) Real-time DNA sequencing using detection of pyrophosphate release. *Anal. Biochem.*, **242**, 84–89.
79. Dressman,D., Yan,H., Traverso,G., Kinzler,K.W. and Vogelstein,B. (2003) Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations. *Proc. Natl Acad. Sci. USA*, **100**, 8817–8822.
80. Bennett,S. (2004) Solexa Ltd. *Pharmacogenomics*, **5**, 433–438.
81. Bennett,S.T., Barnes,C., Cox,A., Davies,L. and Brown,C. (2005) Toward the 1,000 dollars human genome. *Pharmacogenomics*, **6**, 373–382.
82. Goldberg,S.M., Johnson,J., Busam,D., Feldblyum,T., Ferriera,S., Friedman,R., Halpern,A., Khouri,H., Kravitz,S.A. *et al.* (2006) A Sanger/pyrosequencing hybrid approach for the generation of high-quality draft assemblies of marine microbial genomes. *Proc. Natl Acad. Sci. USA*, **103**, 11240–11245.
83. Ng,P., Tan,J.J., Ooi,H.S., Lee,Y.L., Chiu,K.P., Fullwood,M.J., Srinivasan,K.G., Perbost,C., Du,L. *et al.* (2006) Multiplex sequencing of paired-end ditags (MS-PET): a strategy for the ultra-high-throughput analysis of transcriptomes and genomes. *Nucleic Acids Res.*, **34**, e84.
84. Deamer,D.W. and Akeson,M. (2000) Nanopores and nucleic acids: prospects for ultrarapid sequencing. *Trends Biotechnol.*, **18**, 147–151.
85. Guttmacher,A.E. and Collins,F.S. (2002) Genomic medicine–a primer. *N. Engl. J. Med.*, **347**, 1512–1520.
86. Peltonen,L. and McKusick,V.A. (2001) Genomics and medicine. Dissecting human disease in the postgenomic era. *Science*, **291**, 1224–1229.
87. Khoury,M.J., McCabe,L.L. and McCabe,E.R. (2003) Population screening in the age of genomic medicine. *N. Engl. J. Med.*, **348**, 50–58.
88. Drews,J. (2000) Drug discovery: a historical perspective. *Science*, **287**, 1960–1964.
89. Hopkins,A.L. and Groom,C.R. (2002) The druggable genome. *Nat. Rev. Drug Discov.*, **1**, 727–730.
90. Wizemann,T.M., Heinrichs,J.H., Adamou,J.E., Erwin,A.L., Kunsch,C., Choi,G.H., Barash,S.C., Rosen,C.A., Masure,H.R. *et al.* (2001) Use of a whole genome approach to identify vaccine molecules affording protection against *Streptococcus pneumoniae* infection. *Infect. Immun.*, **69**, 1593–1598.
91. Adu-Bobie,J., Capecchi,B., Serruto,D., Rappuoli,R. and Pizza,M. (2003) Two years into reverse vaccinology. *Vaccine*, **21**, 605–610.
92. Pennisi,E. (2006) Genomics. On your mark. Get set. Sequence! *Science*, **314**, 232.
93. Tringe,S.G. and Rubin,E.M. (2005) Metagenomics: DNA sequencing of environmental samples. *Nat. Rev. Genet.*, **6**, 805–814.
94. Tringe,S.G., von Mering,C., Kobayashi,A., Salamov,A.A., Chen,K., Chang,H.W., Podar,M., Short,J.M., Mathur,E.J. *et al.* (2005) Comparative metagenomics of microbial communities. *Science*, **308**, 554–557.
95. Schmidt,T.M., DeLong,E.F. and Pace,N.R. (1991) Analysis of a marine picoplankton community by 16S rRNA gene cloning and sequencing. *J. Bacteriol.*, **173**, 4371–4378.
96. Tyson,G.W., Chapman,J., Hugenholtz,P., Allen,E.E., Ram,R.J., Richardson,P.M., Solovyev,V.V., Rubin,E.M., Rokhsar,D.S. *et al.* (2004) Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature*, **428**, 37–43.
97. Venter,J.C., Remington,K., Heidelberg,J.F., Halpern,A.L., Rusch,D., Eisen,J.A., Wu,D., Paulsen,I., Nelson,K.E. *et al.* (2004) Environmental genome shotgun sequencing of the Sargasso Sea. *Science*, **304**, 66–74.
98. Yutin,N., Suzuki,M.T., Teeling,H., Weber,M., Venter,J.C., Rusch,D.B. and Beja,O. (2007) Assessing diversity and biogeography of aerobic anoxygenic phototrophic bacteria in surface waters of the Atlantic and Pacific Oceans using the Global Ocean Sampling expedition metagenomes. *Environ. Microbiol.*, **9**, 1464–1475.
99. Rusch,D.B., Halpern,A.L., Sutton,G., Heidelberg,K.B., Williamson,S., Yooseph,S., Wu,D., Eisen,J.A., Hoffman,J.M. *et al.* (2007) The Sorcerer II Global Ocean Sampling Expedition: Northwest Atlantic through Eastern Tropical Pacific. *PLoS. Biol.*, **5**, e77.
100. Kannan,N., Taylor,S.S., Zhai,Y., Venter,J.C. and Manning,G. (2007) Structural and Functional Diversity of the Microbial Kinome. *PLoS. Biol.*, **5**, e17.
101. Yooseph,S., Sutton,G., Rusch,D.B., Halpern,A.L., Williamson,S.J., Remington,K., Eisen,J.A., Heidelberg,K.B., Manning,G. *et al.* (2007) The Sorcerer II Global Ocean Sampling Expedition: expanding the Universe of protein families. *PLoS. Biol.*, **5**, e16.
102. Gill,S.R., Pop,M., Deboy,R.T., Eckburg,P.B., Turnbaugh,P.J., Samuel,B.S., Gordon,J.I., Relman,D.A., Fraser-Liggett,C.M. *et al.* (2006) Metagenomic analysis of the human distal gut microbiome. *Science*, **312**, 1355–1359.
103. Zhang,K., Martiny,A.C., Reppas,N.B., Barry,K.W., Malek,J., Chisholm,S.W. and Church,G.M. (2006) Sequencing genomes from single cells by polymerase cloning. *Nat. Biotechnol.*, **24**, 680–686.
104. Hutchison,C.A.III and Venter,J.C. (2006) Single-cell genomics. *Nat. Biotechnol.*, **24**, 657–658.
105. Schaller,R.R. (1997) Moore's law: past, present and future. *IEEE Spectrum*, **34**, 8.