ELSEVIER

Journal of Biotechnology 76 (2000) 1–31

JOURNAL OF Biotechnology

www.elsevier.com/locate/jbiotec

Review article

Sequence analysis of genes and genomes

Fredrik Sterky *, Joakim Lundeberg

Department of Biotechnology, KTH, Royal Institute of Technology, S-100 44 Stockholm, Sweden

Received 15 February 1999; received in revised form 28 May 1999; accepted 25 June 1999

Abstract

A major step towards understanding of the genetic basis of an organism is the complete sequence determination of all genes in its genome. The development of powerful techniques for DNA sequencing has enabled sequencing of large amounts of gene fragments and even complete genomes. Important new techniques for physical mapping, DNA sequencing and sequence analysis have been developed. To increase the throughput, automated procedures for sample preparation and new software for sequence analysis have been applied. This review describes the development of new sequencing methods and the optimisation of sequencing strategies for whole genome and cDNA analysis, as well as discusses issues regarding sequence analysis and annotation.

Keywords: Review; Sequencing technology; Genome sequencing; cDNA sequencing; DNA sequence analysis

1. Historical perspectives

The central dogma of molecular biology describes how genes encoded by DNA sequences are copied (transcribed) to messenger RNA (mRNA), which is then translated into functional proteins. This was first proposed by Francis Crick in 1957 and became the basis of the collinearity hypothesis, which describes that the linear arrangement of subunits in a DNA sequence of a gene corresponds to the amino acid sequence of a protein. By 1966, the entire genetic code was determined (independently by groups of Khorana and Nirenberg), enabling prediction of protein sequences by translation of DNA sequences. Ten years later, robust techniques for rapid DNA sequencing were introduced (Maxam and Gilbert, 1977; Sanger et al., 1977), allowing sequencing of large DNA molecules like the 16.5 kb human mitochondrial genome (Anderson et al., 1981) and the 40 kb genome of the Lambda bacteriophage (Sanger et al., 1982). Since then, the sequencing techniques, and especially the enzymatic chain termination method of Sanger, have been further developed and adapted to different kinds of automation. A dramatic increase in sequence throughput has been accomplished and complete sequencing of large genomes has become possible.

* Corresponding author. Tel.: +46-8-790-8287; fax: +46-8-245-452.

E-mail address: fredrik.sterky@biochem.kth.se (F. Sterky)

0168-1656/00/$ - see front matter
PII: S0168-1656(99)00176-5

This review aims for a broad discussion of the state of the art in DNA sequence analysis with a main focus on whole-genome analysis. It is divided in five major sections, of which the first is a short introduction to the research area. The second section (DNA sequencing technologies) describes the basic techniques for DNA sequencing. The third section (genome sequencing) describes different strategies for mapping and sequencing of whole genomes, while the fourth section (cDNA sequencing) discuss different aspects of cDNA analysis. The last section describes briefly some issues related to analysis and annotation of the generated sequences.

1.1. The genome projects

The human genome project (HGP) was officially launched in 1990 as a 15-year program to map and sequence the entire human genome, an effort that will revolutionise biology and medicine. A number of model organisms representing various forms of life were selected for complete sequencing (Table 1), partly in order to develop new technologies for mapping, sequencing and sequence analysis. In addition, the sequences from these genomes were expected to facilitate elucidation of the functions of genes and sequences in the human genome.

The budding yeast, *Saccharomyces cerevisiae*, was completely sequenced in 1996 (Goffeau et al., 1997) in an international effort and it functions as an excellent model organism (Oliver, 1996). It has a small genome, several chromosomes (16), little repetitive DNA and few introns. It is useful for functional genomics because it is unicellular and can grow on a defined media, it can grow in a haploid or diploid state and tools exist for exact deletion of genes (Rothstein, 1983). There are also striking genetic similarities between human and yeast. Although all human genes have not yet been determined, 31% of the yeast genes have human homologues (Botstein et al., 1997). As much as 70% of the yeast genome consists of coding sequence and only 4.5% of the genes contain introns. Recently the complete genome sequence of the worm *Caenorhabditis elegans* was reported, showing that 36% of the genes have human homologues (The *C. elegans* Sequencing Consortium, 1998). The nematode was chosen as a simple multicellular organism, and it is especially interesting for studies of multicellular development and function of the nervous system (Sulston et al., 1992). *Arabidopsis thaliana* was chosen as a plant model because of its small size, small genome size, rapid growth, low chromosome number and self-fertilisation. It is also low in repetitive DNA and the gene density is one gene per 4.8 kb, as compared to one per 2 kb in yeast. Sequencing of the classical organism for genetic studies, the fruit fly *Drosophila melanogaster*, as well as mouse (*Mus musculus*) is under way. The first free-living organism to be sequenced was the bacteria *Haemophilus influenzae* (Fleischmann et al., 1995) with a genome size of 1.83 Mb, closely followed by *Mycoplasma genitalium* (Fraser et al., 1995), which has the smallest yet published microbial genome (0.58 Mb). To date, 18 microbial genomes have been completely sequenced and published (Table 2), and several others are in progress. In addition, a large number of microbial genomes have been sequenced by

Table 1
Examples of eukaryotic organisms that currently are subjected to whole genome sequencing

Organism	Genome size (Mbp)	No. genes	Sequencing completed	dbEST entries (May 1999)
Homo sapiens	3000	≈80 000	2001*	1 380 000
Mus musculus	3000	≈80 000	2005*	520 000
Drosophila melanogaster	165	12 000	1999*	83 000
Caenorhabditis elegans	100	19 000	1998	73 000
Arabidopsis thaliana	120	21 000	2000*	38 000
Saccharomyces cerevisiae	12	6000	1996	3000

* Estimated time for complete sequence.

Table 2
To date, a total of 18 different bacteria and archaea (a) have been completely sequenced and published. Among the sequencing strategies applied are shotgun sequencing (S), directed sequencing (D) and clone-by clone approaches (C)

Organism	Genome size (Mbp)	Method	Reference
Haemophilus influenzae	1.83	S	Fleischmann et al., 1995
Mycoplasma genitalium	0.58	S	Fraser et al., 1995
Methanococcus jannaschii (a)	1.66	S	Bult et al., 1996
Mycoplasma pneumoniae	0.81	C,D	Himmelreich et al., 1996
Synechocystis sp.	3.57	C,S	Kaneko et al., 1996
Methanobacterium thermoautotrophicum (a)	1.75	S	Smith et al., 1997
Escherichia coli	4.60	C,S	Blattner et al., 1997
Helicobacter pylori (26695)	1.66	S	Tomb et al., 1997
Archaeoglobus fulgidus (a)	2.18	S	Klenk et al., 1997
Borrelia burgdorferi	1.44	S	Fraser et al., 1997
Bacillus subtilis	4.20	C,S,D	Kunst et al., 1997
Mycobacterium tuberculosis	4.40	C,S	Cole et al., 1998
Treponema pallidum	1.14	S	Fraser et al., 1998
Pyrococcus horikoshii (a)	1.80	C,S	Kawarabayasi et al., 1998
Aquifex aeolicus	1.50	S	Deckert et al., 1998
Chlamydia trachomatis	1.04	S	Stephens et al., 1998
Rickettsia prowazekii	1.11	S	Andersson et al., 1998
Helicobacter pylori (J99)	1.64	S	Alm et al., 1999

private initiatives, but these sequences are not publicly available.

In recent years, significant progress has been reached in technologies for genome analysis. For example, optimised cloning vectors for physical mapping, improved enzymes and dye-labels for DNA sequencing, automated sample handling systems, instrumentation for high throughput sequencing and efficient computational tools for analysis of sequence data, have been developed. The sequencing of the human genome has started according to the original plans (Rowen et al., 1997) and will probably be finished in 2003, 2 years ahead of schedule. This plan has been challenged by a private initiative. Celera Genomics, formed by Perkin-Elmer and Craig Venter, will apply a whole genome shotgun approach to complete the human genome sequence in 2001 (Venter et al., 1998).

1.2. Expressed sequence tags (ESTs)

The genomic sequences of higher organisms are only to a minor part represented by protein-coding sequences (3% in humans). However, complementary DNA (cDNA), being reversely transcribed from mRNA, represents a direct source of spliced and coding sequences of genes. Therefore, sequencing of cDNA has become a well-established and accepted technique, which complements the genome sequencing efforts in many important ways.

In 1991, the first application of high-throughput sequencing of cDNA clones was described (Adams et al., 1991), where clones from human brain was randomly selected and partially sequenced. These partial sequences represented genes expressed in the tissue at a certain time-point and were termed ESTs. Similar approaches for different tissues and organisms were soon to follow. The ESTs represent the largest amount of information possible per sequenced base and the vision of rapid identification of all human genes led to the development of several commercially financed data banks with EST sequences. To retain the public interest and competence in this area, Merck Inc. initiated in 1994 the sponsoring of a public EST sequencing effort at Washington University using cDNA clones from the I.M.A.G.E. consortium (Lennon et al., 1996).

EST sequences are extremely valuable for identification of new genes but they are also important

in several other aspects. Sequences obtained by random selection of cDNA clones will yield a statistical picture of the level and complexity of gene expression for the sample tissue. The influence of environmental factors and tissue-specific gene expression can therefore be studied (Okubo et al., 1992). The gene sequences obtained can be efficiently used for physical mapping by determining their chromosomal position. Moreover, they also contribute to the understanding of intron and exon boundaries, which will help to predict the transcribed regions of genomic sequences. These facts confirm the need for cDNA sequencing as a complement to genome sequencing. Furthermore, EST sequencing can be used to reveal single base variations in genes. Detection of these single nucleotide polymorphisms (SNPs) can be an important tool for characterisation of human disease genes (Landegren et al., 1998). Finally, positional cloning of disease genes (Collins, 1995) will be greatly facilitated by the combination of EST sequencing and dense physical maps, and EST sequencing from different organisms will enable evolutionary studies as well as cloning of interesting genes by leaping across taxonomic boundaries (Marra et al., 1998).

2. DNA sequencing technologies

Despite the increasing demand for high throughput in genome sequencing efforts, the basic techniques for DNA sequencing developed more than 20 years ago are still in use. Rather than applying new technologies, an increase in the number of processed samples have been obtained through the development of automated sequencing instruments, robotic sample preparation, optimised chemistry, engineered sequencing enzymes and dyes with higher sensitivity. An important contribution was indeed the development of the polymerase chain reaction (PCR) technique (Saiki et al., 1985), which has revolutionised the ability to detect, analyse and manipulate DNA in many aspects related to genome analysis.

2.1. Labelling and detection principles

The two most important techniques for DNA sequencing are the enzymatic chain termination method (Sanger et al., 1977) and the chemical degradation method (Maxam and Gilbert, 1977). Both are generating a nested set of single-stranded DNA fragments, which are separated by size on an acrylamide gel. As compared to Maxam–Gilbert sequencing, the Sanger chain termination method (Fig. 1) generates more easily interpreted raw data and has become the most widely used. Modifications of the original technique comprise changes in the way of labelling the fragments, development of nucleotides and labels with new chemical properties, automation of signal detection and engineering of enzymes to obtain more reliable raw data.

An important step towards large-scale sequencing was the development of automated DNA sequencers. In these instruments, the electrophoretic step is combined with an on-line detection of fluorescently labelled fragments after excitation by a laser beam. A considerable decrease in time of detection, as well as elimination of the need for radioactive labels, was accomplished. There are two different principles for automated DNA sequencing instruments. Either the fragments are labelled with the same dye in four different reactions (A, C, G and T) and separated in four lanes on the acrylamide gel (Ansorge et al., 1986; Prober et al., 1987; Brumbaugh et al., 1988) or four different dyes are used, enabling the use of only one lane per sample (Smith et al., 1986). The one-dye approach has been commercialised by several manufacturers, for example Amersham Pharmacia Biotech (Automated Laser Fluorescent DNA Sequencer, ALF) and LI-COR (LI-COR DNA Sequencers). The raw data generated is easily interpreted due to the constant mobility of the fragments, which is advantageous in diagnostic and forensic sequencing where slight differences in peak-heights, deriving from polymorphic sites, are analysed. The throughput is lower than on the four-dye systems, although one-dye systems have been adopted to the use of two lasers, enabling the double amount of samples by loading two samples with different

dyes simultaneously (Wiemann et al., 1995). However, if high throughput is desired, the four-dye system is preferable. This approach was first commercialised by Applied Biosystems (ABI 373 and ABI 377) and the ABI instruments have been the choice for most laboratories involved in large scale sequencing, especially since the latest version of the ABI sequencer can analyse up to 96 samples per run. A key parameter with which a system is judged is the accuracy, an issue that is largely dependent on the template and chemistry used (see below). However, considering only the instrumentation, the one-dye systems (ALF, LI-COR) generally yields a higher accuracy than the

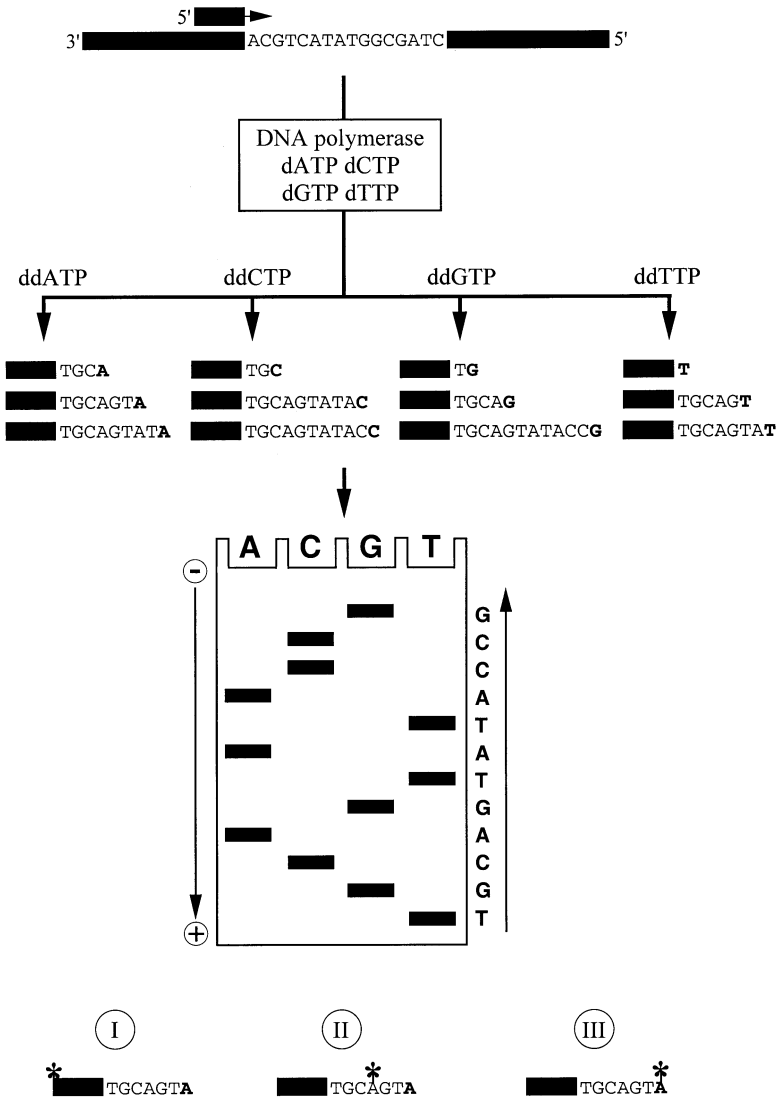


Fig. 1. The Sanger chain termination method. An oligonucleotide primer, hybridised to the template DNA, is used as starting point for synthesis of the complementary strand. By adding a carefully adjusted amount of dideoxy nucleotides, chain termination will occur at positions corresponding to each respective sequence base. The generated nested set of fragments is separated by size on a polyacrylamide gel and the sequence can be determined by the band pattern. The fragments can be radioactively or fluorescently labelled by (I) using a labelled primer, (II) internal labelling or (III) labelled dideoxy nucleotides (dye-terminators).

four-dye instruments (ABI) (Grills et al., 1998). The four-dye systems suffer from the drawback of mobility shifts between the different dyes, which requires the use of more advanced basecalling algorithms to retain accuracy. During recent years, alternative software for basecalling of ABI-traces has been developed. The mostly used, *phred*, is claimed to generate 40–50% fewer errors than the ABI software (Ewing et al., 1998).

There are three different ways of dye labelling of the Sanger fragments (Fig. 1). The dye can be attached to the 5-prime end of the oligonucleotide primer used in the sequencing reaction (Ansorge et al., 1986; Smith et al., 1986) or to the 3-prime end of the fragment (Prober et al., 1987), via a dye-labelled dideoxy nucleotide (dye-terminator). Alternatively, the fragment can be internally labelled by fluorescent nucleotides in an extension/labelling step prior to the chain termination reaction. One advantage with internal labelling and dye-terminators is the flexibility. Any unlabelled primer, synthesised for a certain vector or specific for a cloned template, can be used for sequencing. Furthermore, dye-terminators reduce the number of compressions (unresolved foci in sequencing gels), which is especially important in sequencing of GC-rich regions (Freiberg et al., 1996). The dye-primers are designed for a certain system and are therefore optimal for standardised cloning systems or large scale projects where the same vector is used for a large number of samples.

Much effort has been put into improvements of the time-consuming and expensive gel electrophoresis in Sanger sequencing. Multiplex sequencing (Church and Kieffer-Higgins, 1988) was developed to allow multiple runs simultaneously on one gel. Different vectors were used for the different samples and after separation the fragments were transferred to a membrane, on which consecutive steps of hybridisation and detection were performed with vector-specific oligonucleotides. Although the numbers of gels used decreased dramatically, this method involved additional laborious steps. Alternatively, the speed of electrophoresis could be increased by using thinner gels that allow a higher voltage. Ultrathin gels down to 100 μm have been described for vertical gels (Stegemann et al., 1991) and 25 μm for horizontal gels (Brumley and Smith, 1991). Capillary electrophoresis is another technique that allows higher voltage and decreased separation times. The original methods (Drossman et al., 1990; Luckey et al., 1990; Swerdlow and Gesteland, 1990) used cross-linked polyacrylamide gels as matrices, but the recently described techniques (Carrilho et al., 1996; Goetzinger et al., 1998) use replaceable matrices, which is more cost-effective. Important for the development of capillary electrophoresis instruments was the development of dyes with stronger fluorescent signals (Ju et al., 1995; Metzker et al., 1996; Lee et al., 1997), which enabled detection of smaller amounts of DNA. To date, the slab gel technique has been dominating in large sequencing efforts, but new commercial instruments for capillary electrophoresis like the MegaBACE 1000 DNA Sequencing System (Amersham Pharmacia Biotech/Molecular Dynamics) and ABI 3700 DNA Analyzer (PE Applied Biosystems) indicate an ongoing technology shift.

Among the completely new sequencing technologies described can be mentioned; sequencing by hybridisation (Bains and Smith, 1988; Drmanac et al., 1989; Khrapko et al., 1989; Pease et al., 1994), sequencing by mass spectroscopy (Jacobson et al., 1991; Murray, 1996) and pyrosequencing (Ronaghi et al., 1998). These methods will not be discussed in detail in this review.

2.2. Enzymes for DNA sequencing

Various strategies for DNA sequencing have been described, all representing different combinations of DNA polymerases, dye-labelling techniques and templates. Several different DNA polymerases with different properties have been used, all with special requirements for optimal use. The Klenow fragment (Klenow and Henningsen, 1970), which is a part of *Escherichia coli* DNA polymerase I, was one of the first enzymes used. A drawback with this enzyme is a sequence dependent discrimination of dideoxy nucleotides (ddNTPs) for ordinary nucleotides (dNTPs). This leads to a variation of the amount of fragments for each base in the sequencing reaction and thus, uneven peak-heights in the generated raw data,

which becomes difficult to interpret. The discrimination of dideoxy nucleotides is decreased dramatically by the use of native or modified T7 DNA polymerase (Tabor and Richardson, 1987, 1989) in presence of Mn^{2+} rather than Mg^{2+}. The incorporation of ddNTPs and dNTPs with equal efficiency leads to uniform unambiguous sequence data and higher accuracy after base calling.

Soon after the development of PCR, the thermostable DNA polymerase from *Thermus aquaticus* (*Taq*) (Chien et al., 1976) was used for DNA sequencing. The Sanger fragments are produced by a linear amplification on small amounts of template by the use of temperature cycles as in PCR. This 'cycle sequencing' technique (Innis et al., 1988; Carothers et al., 1989; Murray, 1989) became popular because of its simplicity, but suffered from the same problems as when using the Klenow fragment. The *Taq* DNA polymerase incorporated ddNTPs with poor efficiency (Lee et al., 1992; Khurshid and Beck, 1993), leading to uneven peak-heights and the requirement of large amounts of ddNTPs.

A breakthrough in DNA sequencing came with the discovery that the discrimination that some DNA polymerases showed against ddNTPs was due to a single hydroxyl group on an amino acid in the active sites of the enzymes. Native *E. coli* DNA polymerase I and *Taq* DNA polymerase have a phenylalanine in their active sites, as compared with the tyrosine in T7 DNA polymerase. Exchanging the phenylalanine with a tyrosine in *E. coli* DNA polymerase I and *Taq* DNA polymerase, decreased the discrimination of ddNTPs 250–8000-fold (Tabor and Richardson, 1995). Cycle sequencing with the mutated *Taq* DNA polymerase F667Y resulted in uniform peaks, similar to the pattern obtained by using T7 DNA polymerase. A smaller amount of ddNTPs could be used, which lowered the reagent cost and reduced the fluorescent background in dye-terminator sequencing. Mutants of *Taq* DNA polymerase have had a great impact on DNA sequencing techniques and have been widely used due to their utility in most sequencing applications.

2.3. Template characteristics

The success of a DNA sequencing reaction depends largely on the type and quality of the template. The Sanger sequencing technique is based on the extension of a primer on a single-stranded, or partly single-stranded template. There are a variety of methods available to produce single-stranded DNA (ssDNA) or to manipulate double-stranded DNA (dsDNA) to enable the chain termination reaction.

For sequencing with T7 DNA polymerase or the Klenow fragment, ssDNA can be obtained by cloning the target fragment in M13 phages (Messing et al., 1978). After infection of *E. coli* cells, the excreted single-stranded form of the M13 phage can be collected from the growth medium. This is, however, a labour-intensive task and automated procedures have been developed for large scale sequencing projects (Smith et al., 1990; Zimmermann et al., 1990). An approach to produce single-stranded templates from plasmids is the solid-phase sequencing technique (Ståhl et al., 1988). The technique has been further developed to make use of PCR products as sequencing templates (Hultman et al., 1989). The template is produced by PCR amplification using a biotinylated primer, which enables immobilisation of the PCR product onto streptavidin-coated paramagnetic beads. Strand separation is then obtained by alkali and the beads can be separated from the supernatant by a magnet. Single-stranded template for sequencing is thereby obtained both on the beads and in the supernatant. The solid-phase sequencing technique is suitable for automation (Hultman et al., 1991), which makes it usable for large-scale projects. Several other methods have been developed to convert PCR products into single-stranded templates. Among these are: (1) asymmetric PCR (Gyllensten, 1989); (2) exonuclease generated ssDNA (Higuchi and Ochman, 1989); and (3) transcript sequencing (Sarkar and Sommer, 1988; Stoflet et al., 1988).

Alternatively, double-stranded templates (plasmids or PCR products) can be used directly for sequencing. To make use of plasmids as templates in low-temperature sequencing, they have to be denatured either by heat (Vieira and Messing,

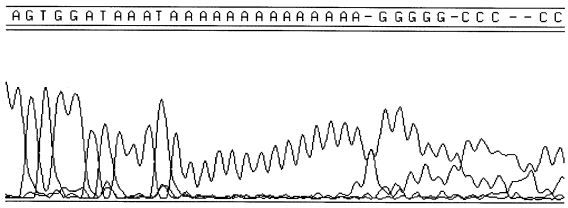


Fig. 2. The results after sequencing of a polymerase chain reaction (PCR) product containing more than 12 consecutive A residues. The *Taq* DNA polymerase fails to incorporate the right number of A:s, which in sequencing leads to that the Sanger fragments goes out of frame.

1982) or by alkali (Chen and Seeburg, 1985). As compared to ssM13 phages, plasmids are easier and quicker to prepare, but these generally yield sequences that are not as long and accurate as for single-stranded templates. In terms of preparation time, PCR products are the best alternative since a template for sequencing can be obtained in a couple of hours by a PCR reaction directly on a bacterial colony or plaque. PCR products have to be treated as plasmids in the low temperature sequencing reactions, but have a tendency to re-associate rapidly, which decreases the primer annealing efficiency. However, heat denaturation followed by snap-cooling in dry-ice ethanol have been successfully used (Kusukawa et al., 1990; Gyllensten et al., 1992).

The requirements for the template used in cycle sequencing are less demanding. As mentioned above, cycle sequencing was developed as an alternative in sequencing of double stranded templates and only small amounts of template is required. Plasmids or PCR products can be used, but also larger constructs like lambda phages and cosmids. The development of modified *Taq* DNA polymerases (Reeve and Fuller, 1995) has resulted in cycle sequencing protocols which produce excellent raw data, leading to that the majority of the high throughput sequencing laboratories uses cycle sequencing protocols.

There are some general differences between M13 phages, plasmids and PCR products as templates in sequencing reactions. The single-stranded M13 phages can only be sequenced on one strand and share a limitation in size with PCR products to a practical maximum of 2–3 kb. Longer PCR products can be obtained under certain conditions (Barnes, 1994; Cheng et al., 1994) but have not been routinely used for sequencing. Plasmids can harbour up to 10 kb fragments if low copy vectors are used. This is a major advantage in primer walking approaches and for gap-filling in shotgun sequencing (see below). A very characteristic property of PCR products in sequencing is the 'polymerase slippage' (Höög, 1991). The term refers to the inability of the *Taq* DNA polymerase to incorporate the correct number of bases in stretches of approximately 12 or more of a certain nucleotide. The polymerase incorporates different number of bases in the stretch region in each cycle, which means that the Sanger fragments will come out of frame in the following sequencing reaction (Fig. 2). This is a serious drawback in attempts to sequence PCR-amplified cDNA clones, which have a poly(A)-tail in the 3-prime end, although anchor primers (5′-oligodT-(A/C/G)-3′) have been used to overcome this problem (Buess et al., 1997).

2.4. High throughput sequencing

Careful considerations in the selection of sequencing technique for high throughput sequencing efforts are required. Sample handling by automated systems is essential and template preparations as well as sequencing reactions have to be optimised for the robotic set-up available. The vision of full integration in a completely automated sequencing system has not yet been reached. More success has been achieved by semi-automated procedures, where each operation is performed separately with a combination of manual and automated procedures. Pipetting of solutions was the first operation to be automated and applied for mixing of sequencing or PCR reactions. Among the commercially available instruments for this, the Biomek 1000 and 2000 (Beckman Instruments) and the Catalyst LabStation (Applied Biosystems) can be mentioned. Operations involving ethanol precipitation are difficult to automate, yet robotics for template preparation of plasmids or phages have been developed. The use of PCR products for sequencing

facilitates this process and has been suggested in the effort to sequence the human genome by the shotgun approach (Venter et al., 1998). In very high throughput sequencing, clone/plaque-picking robots have been used, but loadings of sequencing machines are usually performed manually, a fact that will change with the introduction of new instruments based on capillary electrophoresis. In calculations of cost and throughput, the requirements of skilled researchers to run the automated systems has to be compared with the alternative to use untrained personnel for standardised manual operations.

The sequencing chemistry has to be adapted to automated systems and optimised for high throughput. Cycle sequencing enables sequencing of templates of lower concentration usually produced by the template preparation robots and is the dominating technique, especially after introduction of the new enzymes. Solid-phase technology has been successfully applied to avoid ethanol precipitation in plasmid preparations, sequencing reactions (Hultman et al., 1991) and clean-up of extension products (Tong and Smith, 1992; Sterky et al., 1998). The use of four-colour dye-terminators has enabled performance of sequencing reactions in one tube rather than four tubes. In addition, no labelled primer is required, which is especially important for primer walking strategies.

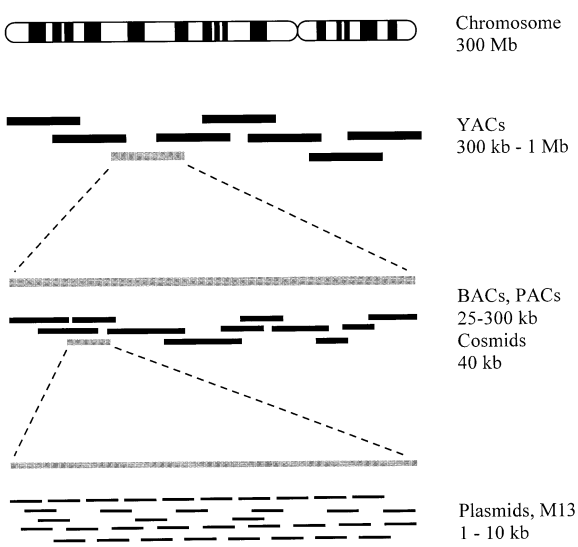


Fig. 3. Examples of vectors used for physical mapping and sequencing and the range of insert sizes that can be harboured by the respective vector.

3. Genome sequencing

There is a dramatic difference in size and complexity between genomes of different organisms. A bacterium can live with a genome as small as 580 kb and about 500 genes (Table 2) while the human genome is more than 5000 times bigger and comprise somewhere between 50 000 and 100 000 genes (Table 1). Furthermore, the organisation of the eukaryotic and prokaryotic genomes is different. The bacterial genomes have a high gene density and lack introns, while eukaryotes contain large regions of non-coding sequences and the genes are in most cases disrupted by introns with the exons scattered over large distances. In addition, eukaryotes contain large regions with repetitive sequences. The strategies for sequencing different genomes have therefore so far been different. However, all genomes have to be subcloned in pieces small enough to suite the sequencing technology and sequence determination should be performed on both strands over the whole genome to ensure sufficient accuracy.

3.1. Physical mapping

Physical mapping is the ordering of clones with large inserts over chromosomal regions. The ordered clones will be the basis for further subcloning into plasmids or phages, which have insert sizes more suitable for the sequencing reactions. The ultimate goal is a physical map that covers the whole genome.

A number of different vectors are available for constructing physical maps, all of which can harbour different sizes of inserts (Fig. 3). Yeast artificial chromosomes (YACs) (Burke et al., 1987) tolerate insert sizes of 250 kb to 1 Mb and have been used to produce a contig (contiguous set of overlapping clones) covering most of the human genome (Cohen et al., 1993). However, the drawback with YACs is a high rate of chimerism and rearrangements (Selleri et al., 1992; Chumakov et

al., 1995). On the other hand, YACs can often hold regions that are difficult to clone in smaller constructs (i.e. cosmids) (Waterston and Sulston, 1995). Cosmids are plasmid vectors that carry the cos sites of the lambda phage (Collins and Hohn, 1978), enabling in vitro packaging in phage particles. The cloned DNA can be efficiently introduced into *E. coli* cells via infection and the cosmid constructs are maintained as plasmids. A cosmid can harbour about 40 kb insert and has been frequently used as a link between YACs and sequencing vectors. Successful use of cosmids to cover large genome regions has been performed (Thierry et al., 1995; Waterston and Sulston, 1995), despite the fact that some regions can be difficult to clone in cosmids and that they are somewhat unstable due to a relatively high copy number. A vector system based on the bacteriophage P1 (Sternberg, 1990) has also been used. The P1 clones could hold inserts up to 100 kb, but the system is a little more complicated to use.

With the human genome in focus, there was a great need for vectors that could stably maintain DNA fragments in a size range between cosmids and YACs. By utilising the *E. coli* F factor, a bacterial artificial chromosome (BAC) was constructed (Shizuya et al., 1992), which could harbour 15–300 kb inserts (usually 100–200 kb). Replication of F vectors are strictly controlled in *E. coli* and the BACs are maintained in a low copy number (one or two copies per cell), thus reducing the potential risk of recombination. The high cloning efficiency by electroporation, easy manipulation and stable maintenance have made BACs more suitable for physical mapping than cosmids. A second system was developed by a modification of the P1 vector. It has similar properties as BACs and the constructs are termed P1-derived artificial chromosomes (PACs) (Ioannou et al., 1994). In recent years, both BACs and PACs have been widely used in the mapping of large genomes.

To create a high-resolution physical map of the human and other genomes, a number of techniques to order the clones must be applied. Similarities in fragment sizes after cleavage with restriction enzymes (fingerprinting) have been used, for example in ordering of λ-clones and cosmids for physical mapping of *C. elegans* (Coulson et al., 1988, 1991), *S. cerevisiae* (Olson et al., 1986; Riles et al., 1993) and *A. thaliana* (Hauge and Goodman, 1992). Cross hybridisation can also be used for ordering of clones, especially to bridge gaps in cosmids contigs with YAC clones (Ward and Jen, 1990). Smaller bridges can alternatively be rescued by long-range PCR. An improvement of the physical mapping procedure came with the use of PCR to amplify unique genomic regions of 100–1000 bases. These sequence tagged sites (STSs) (Olson et al., 1989) only occur in one position in the genome and can therefore be used to order YACs or BACs by PCR screening. STS sequences can be obtained by studying random genomic clones or cDNA clones, or by using polymorphic genetic markers. An STS-based map over the human genome, containing over 30 000 STSs, have been constructed (Hudson et al., 1995; Schuler et al., 1996). This means an average spacing between the STSs of 100 kb, meeting the goal for the human genome project (Collins and Galas, 1993). STS-based maps make it possible to generate extensive physical coverage of a genomic region by screening high-quality BAC libraries.

A physical map is not only a scaffold for genomic sequencing, but also offers the access to any genomic region, which is important for gene cloning. With the knowledge of the genetic location of a disease gene, a collection of overlapping clones encompassing the locus of interest can be obtained, which facilitates isolation of the target gene.

3.2. Random sequencing approaches

There are two major strategies for sequencing of large pieces of DNA, random and directed. The two methods differ in the procedure for cloning, size of inserts and sequencing strategy. In random methods, usually called 'shotgun' sequencing, a library of M13 or plasmid subclones of 1–2 kb inserts is generated. A large number of these clones are then randomly isolated and sequenced using standard vector-specific primers, generating sequences randomly spread over the original fragment (i.e. cosmid or BAC insert). The

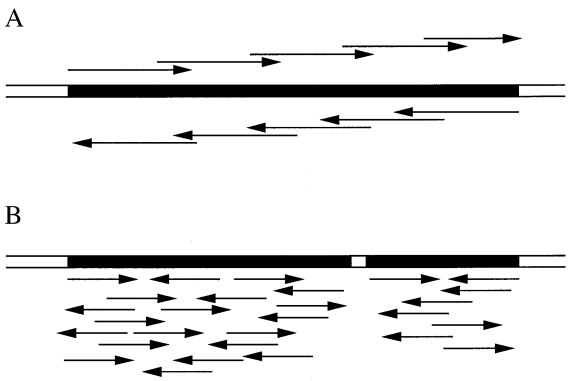


Fig. 4. The principal differences between primer walking and shotgun sequencing. In primer walking (A), sequence data from each reaction are used to design new sequencing primers. A minimal redundancy and full coverage is obtained. In shotgun sequencing (B), randomly selected clones from the insert are sequenced and assembled to form a contiguous sequence. Gaps might have to be filled by directed sequencing methods.

sequence data are then combined and assembled to form contiguous stretches of sequence (contigs) with sequences represented on both strands. Random sequencing is usually performed until 75–95% of the fragment is covered, followed by directed strategies (see below) to fill the gaps.

The fractionation of large DNA fragments into subclones can be obtained by several techniques. Among these are cleavage with restriction enzymes (Messing et al., 1981), treatment with Dnase I (Anderson, 1981), shearing of the DNA by sonication (Deininger, 1983), low pressure (Schriefer et al., 1990), HPLC (Oefner et al., 1996) or 'nebulisation' (GATC GmbH). Restriction enzymes are easy to use for cloning but have several disadvantages. Complete digestion with a single enzyme will produce non-overlapping clones. To overcome this problem, partial digestion or the construction of several sets of clones from different enzymes can be performed to generate sufficient overlap. However, regions with very few restriction sites will be under-represented in the clone banks. The other methods mentioned will break the DNA randomly, producing a set of overlapping fragments. End-repair of the random overhangs produced must be performed, for example using T4 DNA polymerase. A size selection is then usually performed before blunt-end ligation into the sequencing vector. The DNA tends to be more easily broken in AT-rich regions when physically sheared, thus yielding a slight overrepresentation of clones from GC-rich regions. The nebulisation technique was developed to increase the reproducibility and control of the shearing process.

Shotgun sequencing is the traditional method of choice for large-scale sequencing projects. The approach was pioneered in sequencing of the human mitochondrial DNA (Anderson et al., 1981), human adenovirus (Gingeras et al., 1982) and bacteriophage lambda (Sanger et al., 1982). Since then, random shotgun sequencing has been successfully applied in several large-scale DNA sequencing efforts. Typical for shotgun sequencing is the large number of clones that have to be processed to cover the region of interest. Usually, the sequence coverage is 6–8 in a cosmid scale project, which means that every base is represented 6–8 times (redundancy) by different sequence reads. The number of sequencing reactions needed is therefore much higher than for directed strategies, but on the other hand, all reactions can be performed using the same set of primers in a very standardised procedure. The high redundancy also ensures a reliable sequence quality. Shotgun sequencing relies on good computer algorithms to assemble the individual reads into contigs and several such programs are available. To fill the gaps that remain after the shotgun phase, sequencing by primer walking is usually performed on subclones or PCR products bridging the gap (Wilson et al., 1992).

3.3. Directed sequencing approaches

Directed sequencing refers to techniques where the sequence reaction is performed at a known position of the template. The strategy results in a minimal redundancy and a significantly lower number of sequencing reactions as compared to shotgun methods (Fig. 4). A more sophisticated cloning procedure has to be performed for careful selection of template clones.

The most common approach for directed sequencing is primer walking. By this strategy, se-

quence data from one sequence reaction is used to design a new primer for the next reaction (Strauss et al., 1986). A sequence walk can then be performed on both strands of the template. The assembly of contigs is simpler than for shotgun sequencing since the exact position for each sequence reaction is known. The drawback of this methodology is the need for continuous construction of new walking primers. Synthesis of primers is expensive and slow, which reduces the advantage of the lower number of sequence reactions needed. To bypass the traditional way of producing primers, the use of a bank of short oligonucleotides has been suggested. Based on the first generated sequence, two or three adjacent six- to eightmers are selected and ligated on the template (Studier, 1989; Szybalski, 1990) to function as an ordinary primer. The problems arise with incomplete ligation of the 'shortmers' and the large size of a library containing all possible variants. However, hexamers have been reported to function as unique sequencing primers when adjacently annealed without ligation (Kotler et al., 1993) and theoretical considerations of how to reduce the library size by careful selection of shortmers have been published (Blöcker and Lincoln, 1994). Another directed approach that still permits the use of standard sequencing primers is nested deletions. Progressively longer parts of the clone insert are deleted from the vector by different means, bringing more remote regions into the range of a vector-specific sequencing primer. The nested deletions can be obtained by treatment with Exonuclease III (Henikoff, 1984), BAL31 (Misra, 1985) or by the use of internal restriction enzymes.

Directed strategies are often used for closing the remaining gaps after the finished shotgun phase and also for completion of cloned fragments with only partially determined sequence (see Section 4). Entire use of directed approaches in largescale sequencing is still quite rare for several reasons. First, directed strategies require a robust physical map and careful subcloning of the mapped clones (i.e. cosmids or BACs). Secondly, primer synthesis is still expensive despite decreased costs during the last years, and slow for laboratories that do not have access to their own synthesiser. Finally, directed strategies involve more actions from experienced personnel than the standardised procedures in shotgun sequencing. Nevertheless, a few efforts based on directed methods have been published. In the European part of sequencing the whole genome of *S. cerevisiae*, a group at EMBL sequenced the inserts of two cosmids by primer walking with a total redundancy of 2.8 (Wiemann et al., 1993) and 2.6 (Voss et al., 1995). Furthermore, the 0.8 Mb genome of *Mycoplasma pneumoniae* was sequenced by primer walking, nested deletions, a limited shotgun phase and some direct cosmid sequencing (Hilbert et al., 1996), yielding a final redundancy of 2.95 using a total of 5095 sequencing primers. Some efforts have been made to combine the shotgun approach with directed sequencing. In ordered shotgun sequencing (OSS), selection of plasmid templates is made based on overlap of plasmid end-sequences with the growing contig (Chen et al., 1993). Another approach is to create a very detailed physical map of plasmid inserts by hybridisation techniques (Scholler et al., 1995). Standard vector primers can then be used on the selected plasmids to minimise both cost and effort of sequencing.

3.4. How to sequence a genome

The preferable approach for sequencing of large genomes is still under discussion. Obviously, directed strategies can only be considered if a physical map has been or will be constructed. The approach for many sequencing efforts on large eukaryotic genomes, including *S. cerevisiae* (Thierry et al., 1995; Johnston, 1996) and *C. elegans* (Sulston et al., 1992; Wilson et al., 1994) has been construction of overlapping arrays of large insert clones, followed by complete sequencing of these clones one at a time, either by shotgun or directed strategies. Similar strategies have also been applied for smaller bacterial genomes like *M. pneumoniae* (Himmelreich et al., 1996). In these cases, genome sequencing has been initiated after the construction of a physical map.

Shotgun sequencing of whole genomes was first applied at The Institute of Genome Research (TIGR) then headed by Craig Venter. In 1995, the

1.83 Mb genome of *Haemophilus influenzae* was completely determined (Fleischmann et al., 1995) by assembly of 24 000 random sequence reads from plasmids with 1.6–2.0 kb inserts, followed by directed approaches on lambda clones and PCR products to fill the remaining gaps. The overall redundancy was 6.3 and the quality was estimated to 1 error in 5000–10 000 bases. In the very similar strategy for sequencing of the 1.66 Mb genome of *Methanococcus jannaschii* (Bult et al., 1996), greater emphasis was put into end-sequencing of lambda clones with large inserts (16 kb) in parallel to the plasmid sequencing. By including these sequences in the assembly process, the lambda clones were automatically ordered and thus, a physical map was obtained as a result of the assembly. Several other genomes of bacteria and archaea have been sequenced using the whole genome shotgun approach (Table 2).

Sequencing of the human genome has for many years awaited the experiences from the model organisms, but is now under way. The approach taken is the same as for *S. cerevisiae* and *C. elegans*, namely complete sequencing of clone-by-clone selected from physical maps. The available maps, however, are largely based on YACs and have not yet a sufficient resolution (Boguski et al., 1996). High-redundant sequence-ready maps based on BACs, PACs or P1 clones are needed to successfully complete the genome sequence. In 1997, a whole-genome shotgun approach for the human genome was proposed (Weber and Myers, 1997). A combination of small and large plasmids of randomly cloned human DNA from a few individuals was suggested to be sequenced until a 10-fold coverage of the genome (i.e. 30 billion bases) and assembly should preferably be executed by a single large informatics group. The advantages would primarily be the eliminated need for physical mapping, less risk of sequencing recombinant clones since the DNA would be directly cloned in plasmids rather than over other genomic subclones, more polymorphisms detected and most importantly, a lower cost. The proposal was contradicted with the arguments that the shotgun finishing part would be very difficult and expensive (Green, 1997). Arguments for the previously suggested clone-by-clone strategy are advantages like: (1) ability to specifically study problematic regions (modularity); (2) ability to distribute clones to different labs (flexibility); and (3) ability to control the produced sequence by restriction enzymes. Moreover, the existing clones with large inserts enable efficient gap-filling and resequencing of uncertainties. In the whole genome shotgun approach, retracking of the sequenced plasmids would be expensive and the lack of high density physical maps would also make the data less useful for positional cloning of disease genes. Nonetheless, 1 year later, Craig Venter launched the initiation of a whole genome shotgun effort (Venter et al., 1998), aiming to complete the genome in 3 years. The strategy is similar to the one proposed (Weber and Myers, 1997), but with higher emphasis on the use of BACs and BAC-end sequences (Venter et al., 1996) to provide a framework for linking contigs over large regions. It remains to be seen if this strategy will be successful.

4. cDNA sequencing

The ultimate goal for sequencing of whole genomes is to decode all the genetic information carried in the genome. However, there is a drastic variation in the amount of useful information between different organisms. Bacterial genomes, which are small and can effectively be sequenced by whole genome shotgun approaches, also carry compact genetic information with up to 88% coding sequence (*M. genitalium*). In addition, the sequence is easily interpreted due to the lack of introns. Sequencing of larger eukaryotic genomes is less rewarding. The coding sequence in the human genome only represent approximately 3% of the total DNA and for some plant genomes, the figure is even lower. Eukaryotic gene identification by whole genome sequencing is therefore slow and coding regions are difficult to predict due to the introns. A more efficient method for gene identification in eukaryotic genomes, is sequencing of ESTs. ESTs are partial sequences of cDNA, reversely transcribed from mRNA, and represent a direct supply of coding intron-free sequences of genes. During recent years, high

throughput sequencing of ESTs has had a major impact on genome analysis, not only for gene identification, but also for physical mapping and expression profiling.

A major objective of the human genome project is the complete identification of all human genes which of course is the ultimate goal also for other genome projects. EST sequencing has dramatically increased the rate of gene identification and the strategy has been performed for a variety of human tissues (Adams et al., 1991, 1992, 1993, 1995; Gieser and Swaroop, 1992; Khan et al., 1992; Okubo et al., 1992, 1994; Liew, 1993; Takeda et al., 1993; Affara et al., 1994; Liew et al., 1994; Orr et al., 1994; Soares et al., 1994) as well as for mouse (Höög, 1991), *C. elegans* (McCombie et al., 1992; Waterston et al., 1992), *A. thaliana* (Hofte et al., 1993; Newman et al., 1994), rice (Sasaki et al., 1994), maize (Keith et al., 1993), *Brassica napus* (Park et al., 1993), poplar (Sterky et al., 1998) and several other organisms. A special database, dbEST (Boguski et al., 1993), has been established to handle the large amount of ESTs produced. To date (October 1998), dbEST contains 1.8 million entries, 60% of which derives from human tissues.

4.1. mRNA

There are several important differences between prokaryotic and eukaryotic gene expression. Prokaryotic genes are often arranged in operons, which means that several genes are situated after each other on the chromosome and transcribed into one single polycistronic mRNA molecule. Translation into proteins is executed directly without any splicing events. In eukaryotes, each gene is individually translated into a primary RNA transcript, intronic regions are removed by splicing and a mature mRNA transcript is formed. During this process, a CAP structure (7-methyl-GTP) is added to the 5-prime end (Shatkin, 1976) and a poly(A)-tail is added to the 3-prime end of the transcript (Lim and Canellakis, 1970). The mature mRNA is then transported to the cytoplasm and translated to a protein. The gene-coding sequence is flanked by untranslated regions (UTRs) in the mature transcript. The existence of eukaryotic poly(A)-tails is very fortunate since it allows easy purification of mRNA and consequently enables collection of fragments representing all genes expressed in a tissue at a certain time-point.

4.2. Construction of cDNA libraries

A cloning strategy that is well adapted for its purpose is the key to success in cDNA analysis. Different cloning schemes are applied for construction of cDNA libraries optimised for high throughput gene identification by EST sequencing, expression profiling or isolation of full-length clones. A common requirement, however, is that the initial library should be representative, i.e. contain all sequences present in the mRNA population in the same relative frequencies.

Isolation of mRNA usually starts with the preparation of total RNA. In eukaryotic cells, only 1–5% of the total cellular RNA is represented by mRNA, the remainder consisting of ribosomal RNA (rRNA) and transfer RNA (tRNA). All RNA species are very sensitive to degradation by endogenous ribonucleases (RNases) and all labware have to be carefully cleaned before use and the use of RNase inhibitors may be necessary. Rapid snap-freezing of the sample at harvest is also preferable to avoid degradation. RNA can be isolated from subcellular compartments, but usually total RNA is isolated by treatment with guanidinium salts (Chirgwin et al., 1979), followed by phenol–chloroform extraction (Chomczynski and Sacchi, 1987), because of the complete disruption of the cells and rapid inhibition of RNases. Although there is a small but distinct subclass of mRNA which lacks poly(A)-tail (Adesnik et al., 1972; Greenberg and Perry, 1972), most mature mRNAs are tailed with about 40–200 consecutive A:s. Isolation of mRNA from total RNA can therefore be performed by the use of oligo(T)-probes, complementary to the poly(A)-tail. The oligo(T)-probes are usually linked to solid phases like cellulose (Aviv and Leder, 1972), latex particles (Hara et al., 1991) or magnetic beads (Hornes and Korsnes, 1990; Jakobsen et al., 1990). After binding of the mRNA to the solid support, non-

poly(A)-species are washed away and pure mRNA can be eluted, which minimises rRNA contamination.

Conversion of the mRNA to clonable ds-cDNA can be performed in several ways. Synthesis of the first strand of DNA, complementary to the mRNA (i.e. cDNA) is accomplished by the use of reverse transcriptase, primed by oligo(T)-primer which anneals in the poly(A)-tail, random hexamer primers (Dudley et al., 1978) or primers specific for a certain gene (Frohman et al., 1988). For synthesis of the second strand, one of the first methods developed utilises transient hairpin-loops formed in the 3-prime end of the first-strand fragments to prime extension of the second strand (Efstratiadis et al., 1976). The loop structure is removed by S1 nuclease and the ends of the ds-cDNA fragment are polished by T4 DNA polymerase before cloning. Another often used method is a combination of several enzymes (Gubler and Hoffman, 1983); (1) RNase H which nicks the RNA of the RNA-cDNA hybrid; (2) DNA polymerase I which uses the nicked RNA as primer for extension; and (3) DNA ligase which seals remaining nicks in the new DNA strand. Alternatively, a primer site can be introduced in the 3-prime end of the first strand cDNA to enable priming for second strand synthesis. This can be accomplished by homopolymer tailing using terminal transferase (Domec et al., 1990) or ligation of an oligonucleotide to the end of the original mRNA (Kato et al., 1994; Maruyama and Sugano, 1994) or to the end of the cDNA (Apte and Siebert, 1993).

The cDNA is usually cloned in lambda phages or plasmids. Phages tolerate a wider range of insert sizes and therefore yield a more representative library, rich in full length transcripts. Screening of libraries is also easier performed on phage plaques than on bacterial colonies (plasmids). However, plasmids are easier to manipulate, both regarding template preparation and sequencing. Therefore, phage vectors that enables in vitro excision for conversion to plasmids (Short et al., 1988) have been widely used in EST sequencing efforts, although clone representation and frequencies might be altered. In many applications based on cDNA analysis, it is important to know the direction of the clone inserts. Directional cloning can be obtained by including a restriction site (usually *Not*I) to a tail in the oligo(dT)-primer used in the first strand synthesis. After second strand synthesis, an adapter (usually *Eco*RI overhang) is ligated to both ends of the fragment, followed by *Not*I restriction and ligation to the vector. Methylated nucleotides used in the first strand synthesis, efficiently block internal restriction enzyme sites before further cloning. The use of terminal transferase tailing in combination with linkers, has also been described for directional cloning (Han et al., 1987). In contrast to the directed cloning protocols, short fragments randomly distributed along the transcript, can be obtained by using hexamers in the cDNA synthesis (Dudley et al., 1978). Further, in situations when the amount of mRNA is very low (microdissected tissues, etc.), PCR can be used to amplify the cDNA population by using the oligo(dT)-primer tail and the adapter as priming sites (Akowitz and Manuelidis, 1989). Unfortunately, this might lead to an altered representation of original transcripts.

The quality of the cDNA library can be evaluated by different means. Before cloning, the quality of the mRNA population can be determined by northern blot with a probe such as actin, with known size and which is common to most cells. The mRNA homologous to the probe should appear as a distinct band without a large amount of degradation and the total mRNA should be distributed between 0.4 and 4 kb (Moreno-Palanques and Fuldner, 1994). A non-biased representation in the library can be confirmed by plaque screening with three known genes of different expression levels (Okubo et al., 1992). The ratio between the occurrence of positive plaques should be the same as for the signals on a Northern blot for the three genes. A library of good quality, should also contain a minimum of chimeric clones. These can be minimised by using a large excess of adapters and a minor excess of vector during cloning. Long poly(A)-tail can be avoided using a large excess of oligo(dT)-primer in synthesis of the first strand. Since reverse transcriptase has no strand displacement capacity (Moreno-Palanques and Fuldner, 1994), extension can only

occur from the oligo(dT)-primer closest to the 3-prime end of the transcript. Finally, the presence of non-spliced immature mRNA transcripts in cDNA libraries can be minimised by using milder extraction protocols (NP40 detergent) to only recover cytosolic mRNA instead of crude total RNA (Aasheim et al., 1994).

4.3. Gene identification

The strategy for cDNA library construction and the approach for sequencing of the clones are important factors directly influencing the result obtained from the analysis. A summary of cloning and sequencing approaches and their optimal use is shown in Fig. 5. In the initial EST sequencing efforts (Adams et al., 1991, 1992), random primed cDNA libraries were considered to be optimal for gene identification. The problems of sequencing through poly(A)-tails were minimised and the fragments were considered to be enriched in coding sequence rather than the 3-prime or 5-prime UTRs, thus enabling easier functional classification using peptide homology searches. Several other initiatives followed using directed cDNA libraries (Höög, 1991; McCombie et al., 1992;

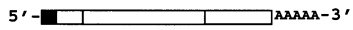
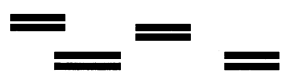
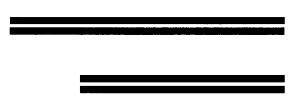
Cloning strategy	Characteristics
mRNA 	The position of the clones in the strategies described are shown in relation to this mRNA molecule.
Random 	The fragments will be distributed over the whole transcript, but the direction of the insert is not known. The inserts often cover coding sequence, which facilitates functional classification by sequence homologies. Problems with poly(A)-tail are minimised, but full-length cDNA clones will not be available.
Full-length 	Libraries optimised to be rich in full-length sequences still contain a fraction of clones that are truncated in the 5-prime end. Coding sequence is therefore often obtained from this end. 3-prime end sequencing using PCR will be difficult due to the poly(A)-tail. Expression analysis by assembly will be incomplete due to non-overlapping sequences from the same transcript.
3-prime end 	Libraries with 3-prime end fragments are best applied in mapping and expression analysis. Sequences from this region will overlap and an assembly will reveal the true expression profile. The 3-prime UTR sequences are poorly conserved and contain few introns, which helps in distinguishing between members of gene families and facilitating the generation of STSs based on these sequences.

Fig. 5. A summary of the alternative cloning strategies for cDNA, and their respective advantages and disadvantages.

Waterston et al., 1992), where the cDNAs were sequenced from the 5-prime end. Gene identification proved to be equally successful as for random clones, mostly because the majority of inserts were truncated in the 5-prime region, thus yielding coding sequence to a large extent. A clear advantage over sequencing of random clones was that the subsequent full-length sequencing of selected genes became much simpler. Sequencing from the 3-prime end has also been performed (Hofte et al., 1993). However, identification of new genes and their functions by homology searches with sequences from the 3-prime end is less rewarding. Further, the success rate is usually lower because of difficulties in sequencing through poly(A)-tails.

The generation of ESTs has also been used for physical mapping (Wilcox et al., 1991; Khan et al., 1992). All ESTs are potential candidates for use as STSs, provided that the sequences are unique and not interrupted by large introns. Since introns are rare in the 3-prime UTRs, sequencing from the 3-prime end is well-suited for mapping purposes. Further, the 3-prime UTR is usually not as conserved as the coding sequence, which facilitates screening of somatic hybrids and distinguishing between members of gene families. As an alternative to sequencing through the poly(A)-tail, special procedures for cloning of the 3-prime region have been developed (Matsubara and Okubo, 1993; Lanfranchi et al., 1996). These methods effectively generate 3-prime sequences that can be used for mapping as well as for expression profiling (see below).

In a typical somatic cell, mRNA species are distributed in three frequency classes (Bishop et al., 1974). There are 10–15 superprevalent genes that represent 10–20% of the total mRNA mass. Intermediate genes (about 1000–2000 genes) represent 40–45% of all mRNAs and rare transcripts (15 000–20 000 genes) represent the remaining 40–45%. Random selection of cDNA clones in EST efforts therefore results in that highly expressed genes are sequenced multiple times. Consequently, effective discovery of new genes is decreasing with the number of clones sequenced from a cDNA library. One way of reducing the sequence redundancy and construct a cDNA library with uniform abundance of transcripts, is normalisation (Ko, 1990; Patanjali et al., 1991; Soares et al., 1994). The methodology is based on reassociation kinetics. The plasmid cDNA library is converted to single-stranded circles and controlled primer extension generates 200 bases long probes. The single-stranded plasmids and the probes are melted apart by heat and reannealed. Rare species will reanneal less rapidly and the single-stranded fraction will become progressively normalised during the reaction. Hydroxyapatite (HAP) columns are then used to bind the double stranded DNA, leaving a single-stranded normalised library in the flow-through. Alternatively, complete elimination (or drastic reduction) of selected species in libraries can be obtained by subtractive hybridisation (Sargent and Dawid, 1983; Sargent, 1987). A library from another tissue can be used to eliminate fragments that occur in both libraries, or a selected number of (frequent) clones can be withdrawn (Bonaldo et al., 1996). If template preparation for sequencing is performed by PCR, non-sequenced clones can be selected by including PCR primers from a small group of highly abundant transcripts (Pacchioni et al., 1996). All PCR products that belong to these transcripts are then detected by occurrence of double bands.

4.4. Expression analysis

Massive EST sequencing from a representative non-normalised library will generate a population of sequences that directly corresponds to the level and complexity of gene expression in the tissue studied. Sequences deriving from the same genes can be grouped into clusters, in which the number of fragments determines the level of expression. Random primed libraries are not suitable for this kind of analysis because non-overlapping sequences from the same gene will frequently occur. Thus, the assembled clusters will not represent the true expression levels. Since the 5-prime end often is truncated in cDNA libraries, 5-prime sequencing yields sequences from the same transcripts at different starting points. Assembly of 5-prime sequences will therefore also yield somewhat false expression levels, but still valuable information

(Rounsley et al., 1996). Alternatively, metabolic differences between tissues can be obtained by grouping the genes in functional categories (Adams et al., 1993; Liew et al., 1994).

Optimal expression profiling is obtained when the sequences from each gene derive from exactly the same place on the gene. Obviously, 3-prime end sequencing would be the best choice but then again, sequencing through the poly(A)-tail is difficult. As an alternative, isolation of the 3-prime region of the cDNA would yield clones that could be sequenced from the 5-prime end but represent the 3-region (Fig. 5). This has been obtained by digestion of the cDNA with frequently cutting (4-cutters) restriction enzymes (Okubo et al., 1992). The drawback is that some transcripts will be lost due to missing enzyme recognition sites, a fact that can be overcome by random fragmentation of the cDNA using sonication (Lanfranchi et al., 1996).

As more and more genes are identified for several organisms, large scale EST sequencing for expression studies become unnecessarily slow and expensive. In 1995, a technique called serial analysis of gene expression (SAGE) was introduced (Velculescu et al., 1995). By a clever cloning strategy involving type IIS restriction enzymes, concatemers of short tags of cDNA could be cloned and sequenced. The sequence tags, not more than 9-mers, were calculated to be unique enough to identify 95% of the human genes. Each sequence run will by this approach generate 30–40 tags, identical for each individual gene. In a following study (Zhang et al., 1997), 300 000 transcripts were used to determine and compare the expression profiles in normal and cancer cells. Another promising tag-sequencing technique, optimal for sequencing from the 3-prime end, is pyrosequencing (Ronaghi et al., 1998). The method represents a completely new non-gel based sequencing technique with potential for very high throughput.

The described EST- and tag-sequencing methods result in complete gene expression profiles. Not only genes that are present or absent in different tissues can be studied, but also up- and down-regulations, which is a great advantage as compared to techniques for differential expression (see below). There are also other methods for complete gene expression analysis. Hybridisation techniques, where RNA or cDNA is hybridised to an array (filter, glass slide or chip) of probes or known cDNA clones have been described (Schena et al., 1995; Drmanac et al., 1996). Yeast is a useful organism for such analysis, where PCR products from most of the 6000 genes can be arrayed, and quantitative expression can be analysed by hybridisation with yeast RNA from different growth conditions (DeRisi et al., 1997). The conventional two-dimensional electrophoresis of proteins also represents an expression profile, but on protein level. This methodology has experienced a renaissance after coupling to methods for protein sequencing by mass spectrometry (Shevchenko et al., 1996).

When time, cost and capacity of sequencing facilities are limited, the difference between two tissues can be studied by identification of the genes that are differentially expressed. These methods do not give an expression profile, but enables isolation of a few candidate genes, responsible for phenotypic or genotypic differences. These methods include differential display (Liang and Pardee, 1992), RNA fingerprinting by arbitrarily primed PCR (Welsh et al., 1992), representational difference analysis (RDA) (Hubank and Schatz, 1994), subtractive hybridisation (Wang and Brown, 1991) and differential screening of arrayed cDNA clones (Byrne et al., 1995).

4.5. Full length sequencing

Determination of the entire protein coding sequence is an important step for functional characterisation of genes. Full-length cDNA sequences are needed for detailed sequence analysis and for the expression of the genes as recombinant proteins. When a partial sequence is known, the initial effort is usually to retrack and sequence the complete insert of the clones from which the selected cDNA sequences are derived. However, the coding sequence might still not be complete. Traditionally, plaque hybridisation has been used to isolate longer fragments from cDNA libraries. This is a time-consuming procedure with a relatively low chance of success, because the phage library and the original library might have been

constructed on the same principles. Even if techniques for enrichment of full length clones have been adapted, many transcripts lack sequence in the 5-prime end due to incomplete first strand synthesis, and even in the 3-prime end due to internal priming or internal restriction sites (i.e. *Not* I).

To specifically determine missing 5-prime or 3-prime sequences, a variety of methods based on the polymerase chain reaction have been developed. Rapid amplification of cDNA ends (RACE) was first described (Frohman et al., 1988), closely followed by the very similar anchored PCR (A-PCR) technique (Loh et al., 1989). To find missing 3-prime ends, the first strand synthesis is performed with an oligo(dT)-primer carrying a PCR primer site in the 5-prime end. Amplification of the candidate gene can then be performed using a primer annealing to the oligo(dT)-tail and a gene-specific primer, designed from the partial cDNA sequence selected (i.e. ESTs or fragments selected by differential approaches). For isolation of the 5-prime end, a gene-specific primer is used for first strand synthesis. A primer site is introduced in the 5-prime end and PCR can be performed with the specific primer and a primer annealing to the new primer site. Originally, the new primer site was introduced by homopolymer tailing using terminal transferase, but greater success have been obtained by using anchor oligonucleotides and T4 RNA ligase (Kato et al., 1994; Maruyama and Sugano, 1994). Biotin capture of the gene-specific first-strand fragments has also been used (Charnock-Jones et al., 1994).

The throughput in full-length sequencing does not match the speed for which new gene candidates are produced by EST efforts or differential expression techniques. Yet, full-length and high quality sequences can potentially provide more accurate database comparisons as well as enhance the ability for different computer programs to predict gene function and structure (Yu et al., 1997). Each selected clone is usually sequenced by primer walking, which is a well established but slow technique. As an alternative, concatenation cDNA sequencing (CCS) has been suggested (Andersson et al., 1997). A selection of cDNA inserts (70 clones) was isolated by restriction enzymes, ligated into concatemers, randomly sheared and cloned in M13 phages. The concatemers were then sequenced by the shotgun approach, enabling simultaneous and accurate sequencing of full-length cDNA clones without the need for walking primers. The generated sequences were electronically cut at the restriction sites prior to assembly and as a result, each cDNA sequence was identified as an individually assembled contig.

5. DNA sequence analysis and annotation

Biological research is now generating sequence data at an explosive rate. The last years, public nucleotide databases have increased with hundreds of million bases per year (Benton, 1996), a rate that will increase further during the scale-up of human genomic sequencing. Management and analysis of the enormous amounts of data will require powerful computational resources. New software and hardware is needed for efficient data processing, assembly and annotation, as well as for gene sequence predictions and functional and structural classifications. Database integration is important to access all kinds of data related to the sequence and the database entries must be automatically updated regularly. Smaller highly annotated data sets for first-pass analyses, will be essential to reduce search times.

5.1. Assembly

Sequence assembly is a process that involves comparison of sequences, finding overlapping fragment pairs, merging as many fragments as possible and creating a consensus sequence from the merged fragments. Accurate assembly algorithms are essential for reconstruction of the original DNA sequence (cosmid, BAC, prokaryotic genome, cDNA clone) in shotgun sequencing and for grouping of ESTs in expression profiling. The challenges for assembly programs are to allow potential sequence ambiguities and still discriminate between repetitive regions, members of gene families or genes sharing the same motif.

There are several assembly algorithms available, including *AutoAssembler* (ABI), *Sequencher*

(Gene Codes Corp.), *GCG* (Dolz, 1994), *GAP4* (Bonfield et al., 1995), *phrap* (Phil Green, weeds.mgh.harvard.edu/goodman/doc/) and the *TIGR Assembler* (Sutton et al., 1995). A commonly used program for cosmid scale assembly is *GAP4*, which is one of several programs in the 'Staden package' (Staden, 1996). It has a highly interactive graphical interface with several tools for manipulation and display of data. The *phrap* assembly program has been successfully used to assemble larger DNA fragments. It works especially well in concert with *phred* (Ewing and Green, 1998), which is an improved lane-tracking and base-calling software, and *consed* (Gordon et al., 1998), which is the graphical interface. The *TIGR Assembler* was developed to manage a whole genome shotgun assembly, i.e. to assemble genomes of 0.5–4 Mb. Such large assemblies means a dramatic increase in the number of pairwise comparisons required, an increased likelihood to encounter repetitive regions and a higher probability to obtain false overlaps due to chimeric clones. The *TIGR Assembler* algorithm has been used for assembly of several prokaryotic genomes as well as for grouping of ESTs for gene indexing purposes.

Assembly algorithms are applied on EST sequences to generate clusters of sequences deriving from the same transcript. It is a way to reduce the large quantity of EST data to a number of groups with overlapping sequences, representing unique genes. Among the efforts to form such gene indices can be mentioned UniGene (Boguski and Schuler, 1995; Schuler et al., 1996), TIGRs human gene index (HGI) (www.tigr.org) and GeneExpress (Houlgatte et al., 1995). The strategies for grouping of the sequences differ in stringency. The *TIGR Assembler* used for HGI, has a stringent matching criterion which prevents chimerism. On the other hand, the strictness results in a more fragmented representation which disallows divergent ESTs that represent alternative forms of the same gene to fold into the same index class. Splice variants are only accepted if they match fully sequenced genes with known isoforms in the Expressed Gene Anatomy Database (EGAD) which is a database with well-characterised human genes (White and Kerlavage, 1996). UniGene and GeneExpress represent looser gene indices (Burke et al., 1998). Sequences are grouped into common classes if they share overlap over a certain threshold, using BLAST, FASTA and Smith–Waterman-methods for comparisons. A single index class can then contain several splice forms of the same gene, but chimeras and other artefacts may be incorrectly included (Houlgatte et al., 1995).

There are several aspects that challenge a successful assembly. First, the error frequency in sequence raw data, which depends on experimental aspects like template, sequencing enzyme, sequencing chemistry and instrumentation, but also on aspects like tracking and base-calling software. A general estimate for the quality of raw data can be obtained from EST sequences, which represent poorly edited, single-pass sequence reads. The overall accuracy for ESTs is usually about 97% (Hillier et al., 1996). The error types include additions, deletions and substitutions of bases and are usually more abundant in the end of a sequence. Other problems that occur are chimeric clones and lane tracking errors and for EST assembly, splice variants, clone reversals and internal priming errors (Gautheret et al., 1998). In the assembly of ESTs, genes belonging to the same gene family can be hard to distinguish. This is especially difficult in plants where as many as 20% (*Arabidopsis*) of the genes belong to gene families (Settles and Byrne, 1998). The higher sequence diversity in the 3-prime UTRs of the transcripts is best used to distinguish between gene family members. Assembly of genomic clones however, meets problems in repetitive regions (SINES, LINES etc.). This is especially difficult for organisms with very high GC-content like *Deinococcus radiodurans* (68% GC) or high AT-content like *Plasmodium falciparum* (82% AT). A correct assembly over long repetitive regions has to be performed with careful respect to the physical map or PCR fragments spanning the region.

5.2. Gene identification

Gene identification and functional classification start with the determination of coding sequence. Basically, an open reading frame (ORF) is deter-

mined from a start codon to a stop codon. This is relatively easy for prokaryotic genomes, where the gene density is high and introns are absent. Usually, ORFs longer than a certain threshold (300–500 bp) are considered as potential genes. Genes that are shorter than the threshold and genes on the opposite strand of longer ORFs (shadow genes) often lead to ambiguities, but can be resolved by analysing the compositional differences between coding regions, shadow genes and non-coding DNA (Burge and Karlin, 1998).

Prediction of coding sequences in eukaryotes is more difficult. The available gene-finding programs generate predictions on the basis of transcriptional signals (transcription start sites, TATA-boxes, polyadenylation sites etc.), translational signals (transcription initiation and termination sites) and splicing signals (donor and acceptor splice site positions). Among these programs are GENEMARK (Borodovsky and McIninch, 1993), GRAIL (Uberbacher et al., 1996) and GENEPARSER (Snyder and Stormo, 1995). These programs are continuously improved, and more advanced programs like GENESCAN (Burge and Karlin, 1997), take into account reading frame compatibility of adjacent exons and compositional properties of introns and exons. Further increase in sensitivity can be obtained by including different sequence similarity functions for comparisons to gene and protein sequences in available databases (Burset and Guigo, 1996). As a complement to the gene identification programs, comparison of complete genomic sequences (20–100 kb) of homologous loci between closely related organisms (mouse and human) can reveal most exons and regulatory regions by identifying regions of particularly high conservation (Hardison et al., 1997).

EST sequences represent spliced genes and are therefore valuable tools for determination of coding sequence in genomic DNA. Comparison between ESTs and genomic sequences immediately reveals the splice sites. However, among the drawbacks are that inconsistencies might occur due to low quality sequence, alternative splicing, presence of pre-mRNA sequences (Wolfsberg and Landsman, 1997) and that the ESTs represent only partial transcript sequences, even after gene-indexing by assembly. Gene annotation techniques based on ESTs (Bailey et al., 1998) and gene-predicting algorithms complement each other in the sense that ESTs are often effective in identifying 3-prime ends of genes where the gene finders often fail, while gene finders relatively well determine the 5-prime ends which the oligo(dT)-primed cDNA clones often fail to reach. This confirms the fact that full understanding of a genome will only be reached by a combination of genomic and cDNA sequencing.

Only a minority of newly sequenced genes have their function determined by an experiment (Andrade and Sander, 1997). Rather, a potential function is proposed by amino acid sequence similarity to earlier characterised proteins, based on the assumption that proteins similar in sequence are also similar in function. Nucleic acid comparisons can also be performed for this purpose, but protein searches are more sensitive because the triplet nucleotide code are degenerate, so the nucleotide sequence can be allowed to change more during evolution but still have the same function. Furthermore, the use of amino acid scoring matrices (PAM, BLOSUM) allows matches between structurally and functionally conserved amino acids, yielding positive scores rather than being regarded as mismatches. The three most widely used algorithms for sequence to sequence comparisons are the basic logical alignment search tool (BLAST) (Altschul et al., 1990), FASTA (Pearson and Lipman, 1988) and the Smith–Waterman algorithm (Smith and Waterman, 1981). A comparison between the original methods reveals differences in terms of speed (BLAST > FASTA $\gg$ Smith–Waterman) and in search performance (Smith–Waterman > FASTA > BLAST) (Pearson, 1995). However, improved functionality has been established in newer versions of BLAST (Altschul et al., 1997) and FASTA (Pearson et al., 1997).

Sequence comparisons are simple and efficient in predicting gene functions. Many identified genes are homologous to genes from completely different organisms. In comparison to the first completely sequenced eukaryote (yeast), other organisms reveal genes with functional similarity in

ranges of 20–25% (bacteria, archaea, human), 25–50% (fungi, plants, protozoans) and over 75% (other yeasts) (Andrade and Sander, 1997). Consequently, a large number of the genes detected in an organism will have homologues in other similar or completely different organisms. For example, comparisons of full length orthologuos mRNA sequences between mouse and human (Makalowski et al., 1996) revealed an average of 85% nucleotide identity in coding sequence and 67–69% identity in 5-prime and 3-prime UTRs. However, sequence similarity does not ensure identical functions, and it is common for groups of genes that are similar in sequence to have diverse (although usually related) functions. Functional predictions can therefore be improved by determining the relation between the sequence to clusters of orthologuos groups (Tatusov et al., 1997) or to evolutionary trees of sequence homologues (Eisen, 1998). The increasing knowledge about gene families and the identification of motifs for gene and protein functional domains will facilitate future annotation of new genes. Further, structural data have opened way for structure/function relationships, which for example is important for rational drug design in the pharmaceutical industry.

5.3. Databases

The genome projects produce large amounts of data and new databases with focus on specific biological areas are constantly developed. The databases provide web-interfaces with different tools for analysis, and in addition, there are several web-sites available that provide powerful search possibilities by alternative algorithms in multiple databases. Despite the great value of increasing amounts of sequences, a few problems will follow: increased search times, increase in high scoring but biologically irrelevant background, inaccurate coding region predictions leading to problems when searching protein databases, and limited first pass annotation (Smith, 1996). To deal with the enormous amount of data, new annotation tools and smaller highly annotated data sets for first-pass analyses will be essential.

Search times may be decreased by reducing the sequence redundancy in the databases as in the gene indexing projects. Further, amino acid data sets are always smaller than nucleotide data sets, especially after the human genome is sequenced, which makes protein searches faster (Anderson and Brass, 1998). Smaller data sets will also decrease the background noise, preventing biologically meaningful similarities to be drowned by random matches. Another problem is the annotation of the sequences. Currently, most protein sequences are translations of gene predictions (Smith, 1996) which may be inaccurate. In addition, annotations get old very quickly and have to be automatically updated or corrected by a third party to keep the database entry updated. It is likely that small, highly annotated data sets with compute-on-demand features, will be available for annotations of new genes in the future.

References

Aasheim, H.C., Deggerdal, A., Smeland, E.B., Hornes, E., 1994. A simple subtraction method for the isolation of cell-specific genes using magnetic monodisperse polymer particles. Biotechniques 16, 716–721.

Adams, M.D., Kelley, J.M., Gocayne, J.D., Dubnick, M., Polymeropoulos, M.H., Xiao, H., Merril, C.R., Wu, A., Olde, B., Moreno, R.F., et al., 1991. Complementary DNA sequencing: expressed sequence tags and human genome project. Science 252, 1651–1656.

Adams, M.D., Dubnick, M., Kerlavage, A.R., Moreno, R., Kelley, J.M., Utterback, T.R., Nagle, J.W., Fields, C., Venter, J.C., 1992. Sequence identification of 2375 human brain genes. Nature 355, 632–634.

Adams, M.D., Kerlavage, A.R., Fields, C., Venter, J.C., 1993. 3400 new expressed sequence tags identify diversity of transcripts in human brain. Nat. Genet. 4, 256–267.

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., White, O., et al., 1995. Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequence. Nature 377, 3–174.

Adesnik, M., Salditt, M., Thomas, W., Darnell, J.E., 1972. Evidence that all messenger RNA molecules except histone messenger RNA contain Poly A sequences and that the Poly A has a nuclear function. J. Mol. Biol. 71, 21–30.

Affara, N.A., Bentley, E., Davey, P., Pelmear, A., Jones, M.H., 1994. The identification of novel gene sequences of the human adult testis. Genomics 22, 205–210.

Akowitz, A., Manuelidis, L., 1989. A novel cDNA/PCR strategy for efficient cloning of small amounts of undefined RNA. Gene 81, 295–306.

Alm, R.A., Ling, L.L., Trust, T.J., et al., 1999. Genomic-sequence comparison of two unrelated isolates of the human gastric pathogen *Helicobacter pylori*. Nature 397, 176–180.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic local alignment search tool. J. Mol. Biol. 215, 403–410.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389–3402.

Anderson, S., 1981. Shotgun DNA sequencing using cloned DNase I-generated fragments. Nucleic Acids Res. 9, 3015–3027.

Anderson, I., Brass, A., 1998. Searching DNA databases for similarities to DNA sequences: when is a match significant? Bioinformatics 14, 349–356.

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., Sanger, F., Schreier, P.H., Smith, A.J., Staden, R., Young, I.G., 1981. Sequence and organization of the human mitochondrial genome. Nature 290, 457–465.

Andersson, B., Lu, J., Shen, Y., Wentland, M.A., Gibbs, R.A., 1997. Simultaneous shotgun sequencing of multiple cDNA clones. DNA Seq. 7, 63–70.

Andersson, S.G., Zomorodipour, A., Andersson, J.O., Sicheritz-Ponten, T., Alsmark, U.C., Podowski, R.M., Naslund, A.K., Eriksson, A.S., Winkler, H.H., Kurland, C.G., 1998. The genome sequence of Rickettsia prowazekii and the origin of mitochondria. Nature 396, 133–140.

Andrade, M.A., Sander, C., 1997. Bioinformatics: from genome data to biological knowledge. Curr. Opin. Biotechnol. 8, 675–683.

Ansorge, W., Sproat, B.S., Stegemann, J., Schwager, C., 1986. A non-radioactive automated method for DNA sequence determination. J. Biochem. Biophys. Methods 13, 315–323.

Apte, A.N., Siebert, P.D., 1993. Anchor-ligated cDNA libraries: a technique for generating a cDNA library for the immediate cloning of the 5′ ends of mRNAs. Biotechniques 15, 890–893.

Aviv, H., Leder, P., 1972. Purification of biologically active globin messenger RNA by chromatography on oligothymidylic acid-cellulose. Proc. Natl. Acad. Sci. USA 69, 1408–1412.

Bailey, L.C. Jr, Searls, D.B., Overton, G.C., 1998. Analysis of EST-driven gene annotation in human genomic sequence. Genome Res. 8, 362–376.

Bains, W., Smith, G.C., 1988. A novel method for nucleic acid sequence determination. J. Theor. Biol. 135, 303–307.

Barnes, W.M., 1994. PCR amplification of up to 35-kb DNA with high fidelity and high yield from lambda bacteriophage templates. Proc. Natl. Acad. Sci. USA 91, 2216–2220.

Benton, D., 1996. Bioinformatics — principles and potential of a new multidisciplinary tool. Trends Biotechnol. 14, 261–272.

Bishop, J.O., Morton, J.G., Rosbash, M., Richardson, M., 1974. Three abundance classes in HeLa cell messenger RNA. Nature 250, 199–204.

Blattner, F.R., Plunkett, G. 3rd, Bloch, C.A., Perna, N.T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J.D., Rode, C.K., Mayhew, G.F., Gregor, J., Davis, N.W., Kirkpatrick, H.A., Goeden, M.A., Rose, D.J., Mau, B., Shao, Y., 1997. The complete genome sequence of Escherichia coli K-12. Science 277, 1453–1474.

Blöcker, H., Lincoln, D.N., 1994. The 'shortmer' approach to nucleic acid sequence analysis. I: Computer simulation of sequencing projects to find economical primer sets. Comput. Appl. Biosci. 10, 193–197.

Boguski, M.S., Schuler, G.D., 1995. ESTablishing a human transcript map. Nat. Genet. 10, 369–371.

Boguski, M.S., Lowe, T.M., Tolstoshev, C.M., 1993. dbEST — database for 'expressed sequence tags'. Nat. Genet. 4, 332–333.

Boguski, M., Chakravarti, A., Gibbs, R., Green, E., Myers, R.M., 1996. The end of the beginning: the race to begin human genome sequencing. Genome Res. 6, 771–772.

Bonaldo, M.F., Lennon, G., Soares, M.B., 1996. Normalization and subtraction: two approaches to facilitate gene discovery. Genome Res. 6, 791–806.

Bonfield, J.K., Smith, K., Staden, R., 1995. A new DNA sequence assembly program. Nucleic Acids Res. 23, 4992–4999.

Borodovsky, M., McIninch, J., 1993. Recognition of genes in DNA sequence with ambiguities. Biosystems 30, 161–171.

Botstein, D., Chervitz, S.A., Cherry, J.M., 1997. Yeast as a model organism. Science 277, 1259–1260.

Brumbaugh, J.A., Middendorf, L.R., Grone, D.L., Ruth, J.L., 1988. Continuous, on-line DNA sequencing using oligodeoxynucleotide primers with multiple fluorophores. Proc. Natl. Acad. Sci. USA 85, 5610–5614.

Brumley, R.L. Jr, Smith, L.M., 1991. Rapid DNA sequencing by horizontal ultrathin gel electrophoresis. Nucleic Acids Res. 19, 4121–4126.

Buess, M., Moroni, C., Hirsch, H.H., 1997. Direct identification of differentially expressed genes by cycle sequencing and cycle labelling using the differential display PCR primers. Nucleic Acids Res. 25, 2233–2235.

Bult, C.J., White, O., Olsen, G.J., Zhou, L., Fleischmann, R.D., Sutton, G.G., Blake, J.A., FitzGerald, L.M., Clayton, R.A., Gocayne, J.D., Kerlavage, A.R., Dougherty, B.A., Tomb, J.F., Adams, M.D., Reich, C.I., Overbeek, R., Kirkness, E.F., Weinstock, K.G., Merrick, J.M., Glodek, A., Scott, J.L., Geoghagen, N.S.M., Venter, J.C., 1996. Complete genome sequence of the methanogenic archaeon, Methanococcus jannaschii. Science 273, 1058–1073.

Burge, C., Karlin, S., 1997. Prediction of complete gene structures in human genomic DNA. J. Mol. Biol. 268, 78–94.

Burge, C.B., Karlin, S., 1998. Finding the genes in genomic DNA. Curr. Opin. Struct. Biol. 8, 346–354.

Burke, D.T., Carle, G.F., Olson, M.V., 1987. Cloning of large segments of exogenous DNA into yeast by means of artificial chromosome vectors. Science 236, 806–812.

Burke, J., Wang, H., Hide, W., Davison, D.B., 1998. Alternative gene form discovery and candidate gene selection from gene indexing projects. Genome Res. 8, 276–290.

Burset, M., Guigo, R., 1996. Evaluation of gene structure prediction programs. Genomics 34, 353–367.

Byrne, J.A., Tomasetto, C., Garnier, J.M., Rouyer, N., Mattei, M.G., Bellocq, J.P., Rio, M.C., Basset, P., 1995. A screening method to identify genes commonly overexpressed in carcinomas and the identification of a novel complementary DNA sequence. Cancer Res. 55, 2896–2903.

Carothers, A.M., Urlaub, G., Mucha, J., Grunberger, D., Chasin, L.A., 1989. Point mutation analysis in a mammalian gene: rapid preparation of total RNA, PCR amplification of cDNA, and Taq sequencing by a novel method. Biotechniques 7, 494–496.

Carrilho, E., Ruiz-Martinez, M.C., Berka, J., Smirnov, I., Goetzinger, W., Miller, A.W., Brady, D., Karger, B.L., 1996. Rapid DNA sequencing of more than 1000 bases per run by capillary electrophoresis using replaceable linear polyacrylamide solutions. Anal. Chem. 68, 3305–3313.

Charnock-Jones, D.S., Platzer, M., Rosenthal, A., 1994. Extension of incomplete cDNAs (ESTs) by biotin/streptavidin-mediated walking using the polymerase chain reaction. J. Biotechnol. 35, 205–215.

Chen, E.Y., Seeburg, P.H., 1985. Supercoil sequencing: a fast and simple method for sequencing plasmid DNA. DNA 4, 165–170.

Chen, E.Y., Schlessinger, D., Kere, J., 1993. Ordered shotgun sequencing, a strategy for integrated mapping and sequencing of YAC clones. Genomics 17, 651–656.

Cheng, S., Fockler, C., Barnes, W.M., Higuchi, R., 1994. Effective amplification of long targets from cloned inserts and human genomic DNA. Proc. Natl. Acad. Sci. USA 91, 5695–5699.

Chien, A., Edgar, D.B., Trela, J.M., 1976. Deoxyribonucleic acid polymerase from the extreme thermophile *Thermus aquaticus*. J. Bacteriol. 127, 1550–1557.

Chirgwin, J.M., Przybyla, A.E., MacDonald, R.J., Rutter, W.J., 1979. Isolation of biologically active ribonucleic acid from sources enriched in ribonuclease. Biochemistry 18, 5294–5299.

Chomczynski, P., Sacchi, N., 1987. Single-step method of RNA isolation by acid guanidinium thiocyanate–phenol–chloroform extraction. Anal. Biochem. 162, 156–159.

Chumakov, I.M., Rigault, P., Le Gall, I., Bellanne-Chantelot, C., Billault, A., Guillou, S., Soularue, P., Guasconi, G., Poullier, E., Gros, I., et al., 1995. A YAC contig map of the human genome. Nature 377, 175–297.

Church, G.M., Kieffer-Higgins, S., 1988. Multiplex DNA sequencing. Science 240, 185–188.

Cohen, D., Chumakov, I., Weissenbach, J., 1993. A first-generation physical map of the human genome. Nature 366, 698–701.

Cole, S.T., Brosch, R., Parkhill, J., Garnier, T., Churcher, C., Harris, D., Gordon, S.V., Eiglmeier, K., Gas, S., Barry, C.E. 3rd, Tekaia, F., Badcock, K., Basham, D., Brown, D., Chillingworth, T., Connor, R., Davies, R., Devlin, K., Feltwell, T., Gentles, S., Hamlin, N., Holroyd, S., Hornsby, T., Jagels, K., Barrell, B.G., et al., 1998. Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence. Nature 393, 537–544.

Collins, F.S., 1995. Positional cloning moves from perditional to traditional. Nat. Genet. 9, 347–350.

Collins, F., Galas, D., 1993. A new five-year plan for the U.S. Human Genome Project. Science 262, 43–46.

Collins, J., Hohn, B., 1978. Cosmids: a type of plasmid gene-cloning vector that is packageable in vitro in bacteriophage lambda heads. Proc. Natl. Acad. Sci. USA 75, 4242–4246.

Coulson, A., Kozono, Y., Lutterbach, B., Shownkeen, R., Sulston, J., Waterston, R., 1991. YACs and the C. elegans genome. Bioessays 13, 413–417.

Coulson, A., Waterston, R., Kiff, J., Sulston, J., Kohara, Y., 1988. Genome linking with yeast artificial chromosomes. Nature 335, 184–186.

Deckert, G., Warren, P.V., Gaasterland, T., Young, W.G., Lenox, A.L., Graham, D.E., Overbeek, R., Snead, M.A., Keller, M., Aujay, M., Huber, R., Feldman, R.A., Short, J.M., Olsen, G.J., Swanson, R.V., 1998. The complete genome of the hyperthermophilic bacterium Aquifex aeolicus. Nature 392, 353–358.

Deininger, P.L., 1983. Random subcloning of sonicated DNA: application to shotgun DNA sequence analysis. Anal. Biochem. 129, 216–223.

DeRisi, J.L., Iyer, V.R., Brown, P.O., 1997. Exploring the metabolic and genetic control of gene expression on a genomic scale. Science 278, 680–686.

Dolz, R., 1994. GCG: fragment assembly programs. Methods Mol. Biol. 24, 9–23.

Domec, C., Garbay, B., Fournier, M., Bonnet, J., 1990. cDNA library construction from small amounts of unfractionated RNA: association of cDNA synthesis with polymerase chain reaction amplification. Anal. Biochem. 188, 422–426.

Drmanac, R., Labat, I., Brukner, I., Crkvenjakov, R., 1989. Sequencing of megabase plus DNA by hybridization: theory of the method. Genomics 4, 114–128.

Drmanac, S., Stavropoulos, N.A., Labat, I., Vonau, J., Hauser, B., Soares, M.B., Drmanac, R., 1996. Gene-representing cDNA clusters defined by hybridization of 57 419 clones from infant brain libraries with short oligonucleotide probes. Genomics 37, 29–40.

Drossman, H., Luckey, J.A., Kostichka, A.J., D'Cunha, J., Smith, L.M., 1990. High-speed separations of DNA sequencing reactions by capillary electrophoresis. Anal. Chem. 62, 900–903.

Dudley, J.P., Butel, J.S., Socher, S.H., Rosen, J.M., 1978. Detection of mouse mammary tumor virus RNA in BALB/c tumor cell lines of nonviral etiologies. J. Virol. 28, 743–752.

Efstratiadis, A., Kafatos, F.C., Maxam, A.M., Maniatis, T., 1976. Enzymatic in vitro synthesis of globin genes. Cell 7, 279–288.

Eisen, J.A., 1998. Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. Genome Res. 8, 163–167.

Ewing, B., Green, P., 1998. Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res. 8, 186–194.

Ewing, B., Hillier, L., Wendl, M.C., Green, P., 1998. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res. 8, 175–185.

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., Merrick, J.M., et al., 1995. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science 269, 496–512.

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., Kelley, J.M., et al., 1995. The minimal gene complement of Mycoplasma genitalium. Science 270, 397–403.

Fraser, C.M., Casjens, S., Huang, W.M., Sutton, G.G., Clayton, R., Lathigra, R., White, O., Ketchum, K.A., Dodson, R., Hickey, E.K., Gwinn, M., Dougherty, B., Tomb, J.F., Fleischmann, R.D., Richardson, D., Peterson, J., Kerlavage, A.R., Quackenbush, J., Salzberg, S., Hanson, M., van Vugt, R., Palmer, N., Adams, M.D., Gocayne, J., Venter, J.C., et al., 1997. Genomic sequence of a Lyme disease spirochaete, Borrelia burgdorferi. Nature 390, 580–586.

Fraser, C.M., Norris, S.J., Weinstock, G.M., White, O., Sutton, G.G., Dodson, R., Gwinn, M., Hickey, E.K., Clayton, R., Ketchum, K.A., Sodergren, E., Hardham, J.M., McLeod, M.P., Salzberg, S., Peterson, J., Khalak, H., Richardson, D., Howell, J.K., Chidambaram, M., Utterback, T., McDonald, L., Artiach, P., Bowman, C., Cotton, M.D., Venter, J.C., et al., 1998. Complete genome sequence of Treponema pallidum, the syphilis spirochete. Science 281, 375–388.

Freiberg, C., Perret, X., Broughton, W.J., Rosenthal, A., 1996. Sequencing the 500-kb GC-rich symbiotic replicon of Rhizobium sp. NGR234 using dye terminators and a thermostable 'sequenase': a beginning. Genome Res. 6, 590–600.

Frohman, M.A., Dush, M.K., Martin, G.R., 1988. Rapid production of full-length cDNAs from rare transcripts: amplification using a single gene-specific oligonucleotide primer. Proc. Natl. Acad. Sci. USA 85, 8998–9002.

Gautheret, D., Poirot, O., Lopez, F., Audic, S., Claverie, J.M., 1998. Alternate polyadenylation in human mRNAs: a large-scale analysis by EST clustering. Genome Res. 8, 524–530.

Gieser, L., Swaroop, A., 1992. Expressed sequence tags and chromosomal localization of cDNA clones from a subtracted retinal pigment epithelium library. Genomics 13, 873–876.

Gingeras, T.R., Sciaky, D., Gelinas, R.E., Bing-Dong, J., Yen, C.E., Kelly, M.M., Bullock, P.A., Parsons, B.L., O'Neill, K.E., Roberts, R.J., 1982. Nucleotide sequences from the adenovirus-2 genome. J. Biol. Chem. 257, 13475–13491.

Goetzinger, W., Kotler, L., Carrilho, E., Ruiz-Martinez, M.C., Salas-Solano, O., Karger, B.L., 1998. Characterization of high molecular mass linear polyacrylamide powder prepared by emulsion polymerization as a replaceable polymer matrix for DNA sequencing by capillary electrophoresis. Electrophoresis 19, 242–248.

Goffeau, A. et al., 1997. The yeast genome directory. Nature 387 (6632 Suppl.).

Gordon, D., Abajian, C., Green, P., 1998. Consed: a graphical tool for sequence finishing. Genome Res. 8, 195–202.

Green, P., 1997. Against a whole-genome shotgun. Genome Res. 7, 410–417.

Greenberg, J.R., Perry, R.P., 1972. Relative occurrence of polyadenylic acid sequences in messenger and heterogeneous nuclear RNA of L cells as determined by poly (U)-hydroxylapatite chromatography. J. Mol. Biol. 72, 91–98.

Grills, G., Dolejsi, M.K., Hardin, S., McMinimy, D., Morrison, P., Rush, J., Scott Adams, P., 1998. Assessing the current state of the art in DNA sequencing. http://abrf.org/ABRF/ResearchCommittees/dsrcreports/DNASEQ98/dsrc98.htm

Gubler, U., Hoffman, B.J., 1983. A simple and very efficient method for generating cDNA libraries. Gene 25, 263–269.

Gyllensten, U.B., 1989. PCR and DNA sequencing. Biotechniques 7, 700–708.

Gyllensten, U.B., Josefsson, A., Schemschat, K., Saldeen, T., Petterson, U., 1992. DNA typing of forensic material with mixed genotypes using allele-specific enzymatic amplification (polymerase chain reaction). Forensic Sci. Int. 52, 149–160.

Han, J.H., Stratowa, C., Rutter, W.J., 1987. Isolation of full-length putative rat lysophospholipase cDNA using improved methods for mRNA isolation and cDNA cloning. Biochemistry 26, 1617–1625.

Hara, E., Kato, T., Nakada, S., Sekiya, S., Oda, K., 1991. Subtractive cDNA cloning using oligo(dT)30-latex and PCR: isolation of cDNA clones specific to undifferentiated human embryonal carcinoma cells. Nucleic Acids Res. 19, 7097–7104.

Hardison, R.C., Oeltjen, J., Miller, W., 1997. Long human-mouse sequence alignments reveal novel regulatory elements: a reason to sequence the mouse genome. Genome Res. 7, 959–966.

Hauge, B.M., Goodman, H.M., 1992. In: Beckmann, J.S., Osborn, T.C. (Eds.), Plant Genomes: Methods for Genetic and Physical Mapping. Kluwer, Dordrecht, The Netherlands, pp. 101–139.

Henikoff, S., 1984. Unidirectional digestion with exonuclease III creates targeted breakpoints for DNA sequencing. Gene 28, 351–359.

Higuchi, R.G., Ochman, H., 1989. Production of single-stranded DNA templates by exonuclease digestion following the polymerase chain reaction. Nucleic Acids Res. 17, 5865.

Hilbert, H., Himmelreich, R., Plagens, H., Herrmann, R., 1996. Sequence analysis of 56 kb from the genome of the bacterium Mycoplasma pneumoniae comprising the dnaA region, the atp operon and a cluster of ribosomal protein genes. Nucleic Acids Res. 24, 628–639.

Hillier, L.D., Lennon, G., Becker, M., Bonaldo, M.F., Chiapelli, B., Chissoe, S., Dietrich, N., DuBuque, T., Favello, A., Gish, W., Hawkins, M., Hultman, M., Kucaba, T., Lacy, M., Le, M., Le, N., Mardis, E., Moore, B., Morris, M., Parsons, J., Prange, C., Rifkin, L., Rohlfing, T., Schellenberg, K., Marra, M., et al., 1996. Generation and analysis of 280 000 human expressed sequence tags. Genome Res. 6, 807–828.

Himmelreich, R., Hilbert, H., Plagens, H., Pirkl, E., Li, B.C., Herrmann, R., 1996. Complete sequence analysis of the genome of the bacterium Mycoplasma pneumoniae. Nucleic Acids Res. 24, 4420–4449.

Hofte, H., Desprez, T., Amselem, J., Chiapello, H., Rouze, P., Caboche, M., Moisan, A., Jourjon, M.F., Charpenteau, J.L., Berthomieu, P., et al., 1993. An inventory of 1152 expressed sequence tags obtained by partial sequencing of cDNAs from Arabidopsis thaliana. Plant. J. 4, 1051–1061.

Höög, C., 1991. Isolation of a large number of novel mammalian genes by a differential cDNA library screening strategy. Nucleic Acids Res. 19, 6123–6127.

Hornes, E., Korsnes, L., 1990. Magnetic DNA hybridization properties of oligonucleotide probes attached to superparamagnetic beads and their use in the isolation of poly(A) mRNA from eukaryotic cells. Genet. Anal. Tech. Appl. 7, 145–150.

Houlgatte, R., Mariage-Samson, R., Duprat, S., Tessier, A., Bentolila, S., Lamy, B., Auffray, C., 1995. The Genexpress Index: a resource for gene discovery and the genic map of the human genome. Genome Res. 5, 272–304.

Hubank, M., Schatz, D.G., 1994. Identifying differences in mRNA expression by representational difference analysis of cDNA. Nucleic Acids Res. 22, 5640–5648.

Hudson, T.J., Stein, L.D., Gerety, S.S., Ma, J., Castle, A.B., Silva, J., Slonim, D.K., Baptista, R., Kruglyak, L., Xu, S.H., et al., 1995. An STS-based map of the human genome. Science 270, 1945–1954.

Hultman, T., Stahl, S., Hornes, E., Uhlen, M., 1989. Direct solid phase sequencing of genomic and plasmid DNA using magnetic beads as solid support. Nucleic Acids Res. 17, 4937–4946.

Hultman, T., Bergh, S., Moks, T., Uhlen, M., 1991. Bidirectional solid-phase sequencing of in vitro-amplified plasmid DNA. Biotechniques 10, 84–93.

Innis, M.A., Myambo, K.B., Gelfand, D.H., Brow, M.A., 1988. DNA sequencing with Thermus aquaticus DNA polymerase and direct sequencing of polymerase chain reaction-amplified DNA. Proc. Natl. Acad. Sci. USA 85, 9436–9440.

Ioannou, P.A., Amemiya, C.T., Garnes, J., Kroisel, P.M., Shizuya, H., Chen, C., Batzer, M.A., de Jong, P.J., 1994. A new bacteriophage P1-derived vector for the propagation of large human DNA fragments. Nat. Genet. 6, 84–89.

Jacobson, K.B., Arlinghaus, H.F., Buchanan, M.V., Chen, C.H., Glish, G.L., Hettich, R.L., McLuckey, S.A., 1991. Applications of mass spectrometry to DNA sequencing. Genet. Anal. Tech. Appl. 8, 223–229.

Jakobsen, K.S., Breivold, E., Hornes, E., 1990. Purification of mRNA directly from crude plant tissues in 15 minutes using magnetic oligo dT microspheres. Nucleic Acids Res. 18, 3669.

Johnston, M., 1996. The complete code for a eukaryotic cell. Genome sequencing. Curr. Biol. 6, 500–503.

Ju, J., Ruan, C., Fuller, C.W., Glazer, A.N., Mathies, R.A., 1995. Fluorescence energy transfer dye-labeled primers for DNA sequencing and analysis. Proc. Natl. Acad. Sci. USA 92, 4347–4351.

Kaneko, T., Sato, S., Kotani, H., Tanaka, A., Asamizu, E., Nakamura, Y., Miyajima, N., Hirosawa, M., Sugiura, M., Sasamoto, S., Kimura, T., Hosouchi, T., Matsuno, A., Muraki, A., Nakazaki, N., Naruo, K., Okumura, S., Shimpo, S., Takeuchi, C., Wada, T., Watanabe, A., Yamada, M., Yasuda, M., Tabata, S., 1996. Sequence analysis of the genome of the unicellular cyanobacterium Synechocystis sp. strain PCC6803. II. Sequence determination of the entire genome and assignment of potential protein-coding regions. DNA Res. 3, 109–136.

Kato, S., Sekine, S., Oh, S.W., Kim, N.S., Umezawa, Y., Abe, N., Yokoyama-Kobayashi, M., Aoki, T., 1994. Construction of a human full-length cDNA bank. Gene 150, 243–250.

Kawarabayasi, Y., Sawada, M., Horikawa, H., Haikawa, Y., Hino, Y., Yamamoto, S., Sekine, M., Baba, S., Kosugi, H., Hosoyama, A., Nagai, Y., Sakai, M., Ogura, K., Otsuka, R., Nakazawa, H., Takamiya, M., Ohfuku, Y., Funahashi, T., Tanaka, T., Kudoh, Y., Yamazaki, J., Kushida, N., Oguchi, A., Aoki, K., Kikuchi, H., 1998. Complete sequence and gene organization of the genome of a hyperthermophilic archaebacterium, Pyrococcus horikoshii OT3. DNA Res. 5, 55–76.

Keith, C.S., Hoang, D.O., Barrett, B.M., Feigelman, B., Nelson, M.C., Thai, H., Baysdorfer, C., 1993. Partial sequence analysis of 130 randomly selected maize cDNA clones. Plant Physiol. 101, 329–332.

Khan, A.S., Wilcox, A.S., Polymeropoulos, M.H., Hopkins, J.A., Stevens, T.J., Robinson, M., Orpana, A.K., Sikela, J.M., 1992. Single pass sequencing and physical and genetic mapping of human brain cDNAs. Nat. Genet. 2, 180–185.

Khrapko, K.R., Lysov Yu, P., Khorlyn, A.A., Shick, V.V., Florentiev, V.L., Mirzabekov, A.D., 1989. An oligonucleotide hybridization approach to DNA sequencing. FEBS Lett. 256, 118–122.

Khurshid, F., Beck, S., 1993. Error analysis in manual and automated DNA sequencing. Anal. Biochem. 208, 138–143.

Klenk, H.P., Clayton, R.A., Tomb, J.F., White, O., Nelson, K.E., Ketchum, K.A., Dodson, R.J., Gwinn, M., Hickey, E.K., Peterson, J.D., Richardson, D.L., Kerlavage, A.R., Graham, D.E., Kyrpides, N.C., Fleischmann, R.D., Quackenbush, J., Lee, N.H., Sutton, G.G., Gill, S., Kirkness, E.F., Dougherty, B.A., McKenney, K., Adams, M.D., Loftus, B., Venter, J.C., et al., 1997. The complete genome sequence of the hyperthermophilic, sulphate-reducing archaeon Archaeoglobus fulgidus. Nature 390, 364–370.

Klenow, H., Henningsen, I., 1970. Selective elimination of the exonuclease activity of the deoxyribonucleic acid polymerase from Escherichia coli B by limited proteolysis. Proc. Natl. Acad. Sci. USA 65, 168–175.

Ko, M.S., 1990. An 'equalized cDNA library' by the reassociation of short double-stranded cDNAs. Nucleic Acids Res. 18, 5705–5711.

Kotler, L.E., Zevin-Sonkin, D., Sobolev, I.A., Beskin, A.D., Ulanovsky, L.E., 1993. DNA sequencing: modular primers assembled from a library of hexamers or pentamers. Proc. Natl. Acad. Sci. USA 90, 4241–4245.

Kunst, F., Ogasawara, N., Moszer, I., Albertini, A.M., Alloni, G., Azevedo, V., Bertero, M.G., Bessieres, P., Bolotin, A., Borchert, S., Borriss, R., Boursier, L., Brans, A., Braun, M., Brignell, S.C., Bron, S., Brouillet, S., Bruschi, C.V., Caldwell, B., Capuano, V., Carter, N.M., Choi, S.K., Codani, J.J., Connerton, I.F., Danchin, A., et al., 1997. The complete genome sequence of the gram-positive bacterium Bacillus subtilis. Nature 390, 249–256.

Kusukawa, N., Uemori, T., Asada, K., Kato, I., 1990. Rapid and reliable protocol for direct sequencing of material amplified by the polymerase chain reaction. Biotechniques 9, 66–68.

Landegren, U., Nilsson, M., Kwok, P.Y., 1998. Reading bits of genetic information: methods for single-nucleotide polymorphism analysis. Genome Res. 8, 769–776.

Lanfranchi, G., Muraro, T., Caldara, F., Pacchioni, B., Pallavicini, A., Pandolfo, D., Toppo, S., Trevisan, S., Scarso, S., Valle, G., 1996. Identification of 4370 expressed sequence tags from a 3′-end-specific cDNA library of human skeletal muscle by DNA sequencing and filter hybridization. Genome Res. 6, 35–42.

Lee, L.G., Connell, C.R., Woo, S.L., Cheng, R.D., McArdle, B.F., Fuller, C.W., Halloran, N.D., Wilson, R.K., 1992. DNA sequencing with dye-labeled terminators and T7 DNA polymerase: effect of dyes and dNTPs on incorporation of dye-terminators and probability analysis of termination fragments. Nucleic Acids Res. 20, 2471–2483.

Lee, L.G., Spurgeon, S.L., Heiner, C.R., Benson, S.C., Rosenblum, B.B., Menchen, S.M., Graham, R.J., Constantinescu, A., Upadhya, K.G., Cassel, J.M., 1997. New energy transfer dyes for DNA sequencing. Nucleic Acids Res. 25, 2816–2822.

Lennon, G., Auffray, C., Polymeropoulos, M., Soares, M.B., 1996. The I.M.A.G.E. Consortium: an integrated molecular analysis of genomes and their expression. Genomics 33, 151–152.

Liang, P., Pardee, A.B., 1992. Differential display of eukaryotic messenger RNA by means of the polymerase chain reaction. Science 257, 967–971.

Liew, C.C., 1993. A human heart cDNA library — the development of an efficient and simple method for automated DNA sequencing. J. Mol. Cell. Cardiol. 25, 891–894.

Liew, C.C., Hwang, D.M., Fung, Y.W., Laurenssen, C., Cukerman, E., Tsui, S., Lee, C.Y., 1994. A catalogue of genes in the cardiovascular system as identified by expressed sequence tags. Proc. Natl. Acad. Sci. USA 91, 10645–10649.

Lim, L., Canellakis, E.S., 1970. Adenine-rich polymer associated with rabbit reticulocyte messenger RNA. Nature 227, 710–712.

Loh, E.Y., Elliott, J.F., Cwirla, S., Lanier, L.L., Davis, M.M., 1989. Polymerase chain reaction with single-sided specificity: analysis of T cell receptor delta chain. Science 243, 217–220.

Luckey, J.A., Drossman, H., Kostichka, A.J., Mead, D.A., D'Cunha, J., Norris, T.B., Smith, L.M., 1990. High speed DNA sequencing by capillary electrophoresis. Nucleic Acids Res. 18, 4417–4421.

Makalowski, W., Zhang, J., Boguski, M.S., 1996. Comparative analysis of 1196 orthologous mouse and human full-length mRNA and protein sequences. Genome Res. 6, 846–857.

Marra, M.A., Hillier, L., Waterston, R.H., 1998. Expressed sequence tags — ESTablishing bridges between genomes. Trends Genet. 14, 4–7.

Maruyama, K., Sugano, S., 1994. Oligo-capping: a simple method to replace the cap structure of eukaryotic mRNAs with oligoribonucleotides. Gene 138, 171–174.

Matsubara, K., Okubo, K., 1993. cDNA analyses in the human genome project. Gene 135, 265–274.

Maxam, A.M., Gilbert, W., 1977. A new method for sequencing DNA. Proc. Natl. Acad. Sci. USA 74, 560–564.

McCombie, W.R., Adams, M.D., Kelley, J.M., FitzGerald, M.G., Utterback, T.R., Khan, M., Dubnick, M., Kerlavage, A.R., Venter, J.C., Fields, C., 1992. Caenorhabditis elegans expressed sequence tags identify gene families and potential disease gene homologues. Nat. Genet. 1, 124–131.

Messing, J., Gronenborn, B., Muller-Hill, B., Hofschneider, P.H., 1978. Single-stranded filamentous DNA phage as a carrier for in vitro recombined DNA. In: Hofschneider, P.H., Starlinger, P. (Eds.), Integration and Excision of DNA molecules, vol. 26. Springer, Berlin, pp. 29–32.

Messing, J., Crea, R., Seeburg, P.H., 1981. A system for shotgun DNA sequencing. Nucleic Acids Res. 9, 309–321.

Metzker, M.L., Lu, J., Gibbs, R.A., 1996. Electrophoretically uniform fluorescent dyes for automated DNA sequencing. Science 271, 1420–1422.

Misra, T.K., 1985. A new strategy to create ordered deletions for rapid nucleotide sequencing. Gene 34, 263–268.

Moreno-Palanques, R.F., Fuldner, R.A., 1994. In: Adams, M.D., Fields, C., Venter, J.C. (Eds.), Automated DNA Sequencing: Construction of cDNA Libraries. Academic Press, London, pp. 102–109.

Murray, V., 1989. Improved double-stranded DNA sequencing using the linear polymerase chain reaction. Nucleic Acids Res. 17, 8889.

Murray, K.K., 1996. DNA sequencing by mass spectrometry. J. Mass Spectrom. 31, 1203–1215.

Newman, T., de Bruijn, F.J., Green, P., Keegstra, K., Kende, H., McIntosh, L., Ohlrogge, J., Raikhel, N., Somerville, S., Thomashow, M., et al., 1994. Genes galore: a summary of methods for accessing results from large-scale partial sequencing of anonymous Arabidopsis cDNA clones. Plant. Physiol. 106, 1241–1255.

Oefner, P.J., Hunicke-Smith, S.P., Chiang, L., Dietrich, F., Mulligan, J., Davis, R.W., 1996. Efficient random subcloning of DNA sheared in a recirculating point-sink flow system. Nucleic Acids Res. 24, 3879–3886.

Okubo, K., Hori, N., Matoba, R., Niiyama, T., Fukushima, A., Kojima, Y., Matsubara, K., 1992. Large scale cDNA sequencing for analysis of quantitative and qualitative aspects of gene expression. Nat. Genet. 2, 173–179.

Okubo, K., Yoshii, J., Yokouchi, H., Kameyama, M., Matsubara, K., 1994. An expression profile of active genes in human colonic mucosa. DNA Res. 1, 37–45.

Oliver, S., 1996. A network approach to the systematic analysis of yeast gene function. Trends Genet. 12, 241–242.

Olson, M.V., Dutchik, J.E., Graham, M.Y., Brodeur, G.M., Helms, C., Frank, M., MacCollin, M., Scheinman, R., Frank, T., 1986. Random-clone strategy for genomic restriction mapping in yeast. Proc. Natl. Acad. Sci. USA 83, 7826–7830.

Olson, M., Hood, L., Cantor, C., Botstein, D., 1989. A common language for physical mapping of the human genome. Science 245, 1434–1435.

Orr, S.L., Hughes, T.P., Sawyers, C.L., Kato, R.M., Quan, S.G., Williams, S.P., Witte, O.N., Hood, L., 1994. Isolation of unknown genes from human bone marrow by differential screening and single-pass cDNA sequence determination. Proc. Natl. Acad. Sci. USA 91, 11869–11873.

Pacchioni, B., Trevisan, S., Gomirato, S., Toppo, S., Valle, G., Lanfranchi, G., 1996. Semi-multiplex PCR technique for screening of abundant transcripts during systematic sequencing of cDNA libraries. Biotechniques 21, 644–646.

Park, Y.S., Kwak, J.M., Kwon, O.Y., Kim, Y.S., Lee, D.S., Cho, M.J., Lee, H.H., Nam, H.G., 1993. Generation of expressed sequence tags of random root cDNA clones of Brassica napus by single-run partial sequencing. Plant Physiol. 103, 359–370.

Patanjali, S.R., Parimoo, S., Weissman, S.M., 1991. Construction of a uniform-abundance (normalized) cDNA library. Proc. Natl. Acad. Sci. USA 88, 1943–1947.

Pearson, W.R., 1995. Comparison of methods for searching protein sequence databases. Protein Sci. 4, 1145–1160.

Pearson, W.R., Lipman, D.J., 1988. Improved tools for biological sequence comparison. Proc. Natl. Acad. Sci. USA 85, 2444–2448.

Pearson, W.R., Wood, T., Zhang, Z., Miller, W., 1997. Comparison of DNA sequences with protein sequences. Genomics 46, 24–36.

Pease, A.C., Solas, D., Sullivan, E.J., Cronin, M.T., Holmes, C.P., Fodor, S.P., 1994. Light-generated oligonucleotide arrays for rapid DNA sequence analysis. Proc. Natl. Acad. Sci. USA 91, 5022–5026.

Prober, J.M., Trainor, G.L., Dam, R.J., Hobbs, F.W., Robertson, C.W., Zagursky, R.J., Cocuzza, A.J., Jensen, M.A., Baumeister, K., 1987. A system for rapid DNA sequencing with fluorescent chain-terminating dideoxynucleotides. Science 238, 336–341.

Reeve, M.A., Fuller, C.W., 1995. A novel thermostable polymerase for DNA sequencing. Nature 376, 796–797.

Riles, L., Dutchik, J.E., Baktha, A., McCauley, B.K., Thayer, E.C., Leckie, M.P., Braden, V.V., Depke, J.E., Olson, M.V., 1993. Physical maps of the six smallest chromosomes of Saccharomyces cerevisiae at a resolution of 2.6 kilobase pairs. Genetics 134, 81–150.

Ronaghi, M., Uhlen, M., Nyren, P., 1998. A sequencing method based on real-time pyrophosphate. Science 281, 363.

Rothstein, R.J., 1983. One-step gene disruption in yeast. Methods Enzymol. 101, 202–211.

Rounsley, S.D., Glodek, A., Sutton, G., Adams, M.D., Somerville, C.R., Venter, J.C., Kerlavage, A.R., 1996. The construction of Arabidopsis expressed sequence tag assemblies. A new resource to facilitate gene identification. Plant Physiol. 112, 1177–1183.

Rowen, L., Mahairas, G., Hood, L., 1997. Sequencing the human genome. Science 278, 605–607.

Saiki, R.K., Scharf, S., Faloona, F., Mullis, K.B., Horn, G.T., Erlich, H.A., Arnheim, N., 1985. Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. Science 230, 1350–1354.

Sanger, F., Nicklen, S., Coulson, A.R., 1977. DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. USA 74, 5463–5467.

Sanger, F., Coulson, A.R., Hong, G.F., Hill, D.F., Petersen, G.B., 1982. Nucleotide sequence of bacteriophage lambda DNA. J. Mol. Biol. 162, 729–773.

Sargent, T.D., 1987. Isolation of differentially expressed genes. Methods Enzymol. 152, 423–432.

Sargent, T.D., Dawid, I.B., 1983. Differential gene expression in the gastrula of Xenopus laevis. Science 222, 135–139.

Sarkar, G., Sommer, S.S., 1988. RNA amplification with transcript sequencing (RAWTS). Nucleic Acids Res. 16, 5197.

Sasaki, T., Song, J., Koga-Ban, Y., Matsui, E., Fang, F., Higo, H., Nagasaki, H., Hori, M., Miya, M., Murayama-Kayano, E., et al., 1994. Toward cataloguing all rice genes: large-scale sequencing of randomly chosen rice cDNAs from a callus cDNA library. Plant J. 6, 615–624.

Schena, M., Shalon, D., Davis, R.W., Brown, P.O., 1995. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science 270, 467–470.

Scholler, P., Karger, A.E., Meier-Ewert, S., Lehrach, H., Delius, H., Hoheisel, J.D., 1995. Fine-mapping of shotgun template-libraries; an efficient strategy for the systematic sequencing of genomic DNA. Nucleic Acids Res. 23, 3842–3849.

Schriefer, L.A., Gebauer, B.K., Qui, L.Q., Waterston, R.H., Wilson, R.K., 1990. Low pressure DNA shearing: a method for random DNA sequence analysis. Nucleic Acids Res. 18, 7455–7456.

Schuler, G.D., Boguski, M.S., Stewart, E.A., Stein, L.D., Gyapay, G., Rice, K., White, R.E., Rodriguez-Tome, P., Aggarwal, A., Bajorek, E., Bentolila, S., Birren, B.B., Butler, A., Castle, A.B., Chiannilkulchai, N., Chu, A., Clee, C., Cowles, S., Day, P.J., Dibling, T., Drouot, N., Dunham, I., Duprat, S., East, C., Hudson, T.J., et al., 1996. A gene map of the human genome. Science 274, 540–546.

Selleri, L., Eubanks, J.H., Giovannini, M., Hermanson, G.G., Romo, A., Djabali, M., Maurer, S., McElligott, D.L., Smith, M.W., Evans, G.A., 1992. Detection and characterization of 'chimeric' yeast artificial chromosome clones by fluorescent in situ suppression hybridization. Genomics 14, 536–541.

Settles, A.M., Byrne, M., 1998. Opportunities and challenges grow from Arabidopsis genome sequencing. Genome Res. 8, 83–85.

Shatkin, A.J., 1976. Capping of eucaryotic mRNAs. Cell 9, 645–653.

Shevchenko, A., Jensen, O.N., Podtelejnikov, A.V., Sagliocco, F., Wilm, M., Vorm, O., Mortensen, P., Boucherie, H., Mann, M., 1996. Linking genome and proteome by mass spectrometry: large-scale identification of yeast proteins from two dimensional gels. Proc. Natl. Acad. Sci. USA 93, 14440–14445.

Shizuya, H., Birren, B., Kim, U.J., Mancino, V., Slepak, T., Tachiiri, Y., Simon, M., 1992. Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in Escherichia coli using an F-factor-based vector. Proc. Natl. Acad. Sci. USA 89, 8794–8797.

Short, J.M., Fernandez, J.M., Sorge, J.A., Huse, W.D., 1988. Lambda ZAP: a bacteriophage lambda expression vector with in vivo excision properties. Nucleic Acids Res. 16, 7583–7600.

Smith, R.F., 1996. Perspectives: sequence data base searching in the era of large-scale genomic sequencing. Genome Res. 6, 653–660.

Smith, L.M., Sanders, J.Z., Kaiser, R.J., Hughes, P., Dodd, C., Connell, C.R., Heiner, C., Kent, S.B., Hood, L.E., 1986. Fluorescence detection in automated DNA sequence analysis. Nature 321, 674–679.

Smith, V., Brown, C.M., Bankier, A.T., Barrell, B.G., 1990. Semiautomated preparation of DNA templates for large-scale sequencing projects. DNA Seq. 1, 73–78.

Smith, D.R., Doucette-Stamm, L.A., Deloughery, C., Lee, H., Dubois, J., Aldredge, T., Bashirzadeh, R., Blakely, D., Cook, R., Gilbert, K., Harrison, D., Hoang, L., Keagle, P., Lumm, W., Pothier, B., Qiu, D., Spadafora, R., Vicaire, R., Wang, Y., Wierzbowski, J., Gibson, R., Jiwani, N., Caruso, A., Bush, D., Reeve, J.N., et al., 1997. Complete genome sequence of Methanobacterium thermoautotrophicum deltaH: functional analysis and comparative genomics. J. Bacteriol. 179, 7135–7155.

Smith, T.F., Waterman, M.S., 1981. Identification of common molecular subsequences. J. Mol. Biol. 147, 195–197.

Snyder, E.E., Stormo, G.D., 1995. Identification of protein coding regions in genomic DNA. J. Mol. Biol. 248, 1–18.

Soares, M.B., Bonaldo, M.F., Jelene, P., Su, L., Lawton, L., Efstratiadis, A., 1994. Construction and characterization of a normalized cDNA library. Proc. Natl. Acad. Sci. USA 91, 9228–9232.

Staden, R., 1996. The Staden sequence analysis package. Mol. Biotechnol. 5, 233–241.

Ståhl, S., Hultman, T., Olsson, A., Moks, T., Uhlen, M., 1988. Solid phase DNA sequencing using the biotin-avidin system. Nucleic Acids Res. 16, 3025–3038.

Stegemann, J., Schwager, C., Erfle, H., Hewitt, N., Voss, H., Zimmermann, J., Ansorge, W., 1991. High speed on-line DNA sequencing on ultrathin slab gels. Nucleic Acids Res. 19, 675–676.

Stephens, R.S., Kalman, S., Lammel, C., Fan, J., Marathe, R., Aravind, L., Mitchell, W., Olinger, L., Tatusov, R.L., Zhao, Q., Koonin, E.V., Davis, R.W., 1998. Genome sequence of an obligate intracellular pathogen of humans: Chlamydia trachomatis. Science 282, 754–759.

Sterky, F., Regan, S., Karlsson, J., Hertzberg, M., Rohde, A., Holmberg, A., Amini, B., Bhalerao, R., Larsson, M., Villarroel, R., Van Montagu, M., Sandberg, G., Olsson, O., Teeri, T.T., Boerjan, W., Gustafsson, P., Uhlen, M., Sundberg, B., Lundeberg, J., 1998. Gene discovery in the wood-forming tissues of poplar: analysis of 5692 expressed sequence tags. Proc. Natl. Acad. Sci. USA 95, 13330–13335.

Sternberg, N., 1990. Bacteriophage P1 cloning system for the isolation, amplification, and recovery of DNA fragments as large as 100 kilobase pairs. Proc. Natl. Acad. Sci. USA 87, 103–107.

Stoflet, E.S., Koeberl, D.D., Sarkar, G., Sommer, S.S., 1988. Genomic amplification with transcript sequencing. Science 239, 491–494.

Strauss, E.C., Kobori, J.A., Siu, G., Hood, L.E., 1986. Specific-primer-directed DNA sequencing. Anal. Biochem. 154, 353–360.

Studier, F.W., 1989. A strategy for high-volume sequencing of cosmid DNAs: random and directed priming with a library of oligonucleotides. Proc. Natl. Acad. Sci. USA 86, 6917–6921.

Sulston, J., Du, Z., Thomas, K., Wilson, R., Hillier, L., Staden, R., Halloran, N., Green, P., Thierry-Mieg, J., Qiu, L., et al., 1992. The C. elegans genome sequencing project: a beginning. Nature 356, 37–41.

Sutton, G.G., White, O., Adams, M.D., Kerlavage, A.R., 1995. TIGR Assembler: A new tool for assembling large shotgun sequencing projects. Genome Sci. Tech. 1, 9–19.

Swerdlow, H., Gesteland, R., 1990. Capillary gel electrophoresis for rapid, high resolution DNA sequencing. Nucleic Acids Res. 18, 1415–1419.

Szybalski, W., 1990. Proposal for sequencing DNA using ligation of hexamers to generate sequential elongation primers (SPEL-6). Gene 90, 177–178.

Tabor, S., Richardson, C.C., 1987. DNA sequence analysis with a modified bacteriophage T7 DNA polymerase. Proc. Natl. Acad. Sci. USA 84, 4767–4771.

Tabor, S., Richardson, C.C., 1989. Selective inactivation of the exonuclease activity of bacteriophage T7 DNA polymerase by in vitro mutagenesis. J. Biol. Chem. 264, 6447–6458.

Tabor, S., Richardson, C.C., 1995. A single residue in DNA polymerases of the Escherichia coli DNA polymerase I family is critical for distinguishing between deoxy- and dideoxyribonucleotides. Proc. Natl. Acad. Sci. USA 92, 6339–6343.

Takeda, J., Yano, H., Eng, S., Zeng, Y., Bell, G.I., 1993. A molecular inventory of human pancreatic islets: sequence analysis of 1000 cDNA clones. Hum. Mol. Genet. 2, 1793–1798.

Tatusov, R.L., Koonin, E.V., Lipman, D.J., 1997. A genomic perspective on protein families. Science 278, 631–637.

The C. elegans Sequencing Consortium, 1998. Genome sequence of the nematode C. elegans: a platform for investigating biology. The C. elegans Sequencing Consortium. Science 282, 2012–2018.

Thierry, A., Gaillon, L., Galibert, F., Dujon, B., 1995. Construction of a complete genomic library of Saccharomyces cerevisiae and physical mapping of chromosome XI at 3.7 kb resolution. Yeast 11, 121–135.

Tomb, J.F., White, O., Kerlavage, A.R., Clayton, R.A., Sutton, G.G., Fleischmann, R.D., Ketchum, K.A., Klenk, H.P., Gill, S., Dougherty, B.A., Nelson, K., Quackenbush, J., Zhou, L., Kirkness, E.F., Peterson, S., Loftus, B., Richardson, D., Dodson, R., Khalak, H.G., Glodek, A., McKenney, K., Fitzegerald, L.M., Lee, N., Adams, M.D., Venter, J.C., et al., 1997. The complete genome sequence of the gastric pathogen Helicobacter pylori. Nature 388, 539–547.

Tong, X., Smith, L.M., 1992. Solid-phase method for the purification of DNA sequencing reactions. Anal. Chem. 64, 2672–2677.

Uberbacher, E.C., Xu, Y., Mural, R.J., 1996. Discovering and understanding genes in human DNA sequence using GRAIL. Methods Enzymol. 266, 259–281.

Velculescu, V.E., Zhang, L., Vogelstein, B., Kinzler, K.W., 1995. Serial analysis of gene expression. Science 270, 484–487.

Venter, J.C., Smith, H.O., Hood, L., 1996. A new strategy for genome sequencing. Nature 381, 364–366.

Venter, J.C., Adams, M.D., Sutton, G.G., Kerlavage, A.R., Smith, H.O., Hunkapiller, M., 1998. Shotgun sequencing of the human genome. Science 280, 1540–1542.

Vieira, J., Messing, J., 1982. The pUC plasmids, an M13mp7-derived system for insertion mutagenesis and sequencing with synthetic universal primers. Gene 19, 259–268.

Voss, H., Schwager, C., Wiemann, S., Zimmermann, J., Stegemann, J., Erfle, H., Voie, A.M., Drzonek, H., Ansorge, W., 1995. Efficient low redundancy large-scale DNA sequencing at EMBL. J. Biotechnol. 41, 121–129.

Wang, Z., Brown, D.D., 1991. A gene expression screen. Proc. Natl. Acad. Sci. USA 88, 11505–11509.

Ward, E.R., Jen, G.C., 1990. Isolation of single-copy-sequence clones from a yeast artificial chromosome library of randomly-sheared Arabidopsis thaliana DNA. Plant. Mol. Biol. 14, 561–568.

Waterston, R., Sulston, J., 1995. The genome of Caenorhabditis elegans. Proc. Natl. Acad. Sci. USA 92, 10836–10840.

Waterston, R., Martin, C., Craxton, M., Huynh, C., Coulson, A., Hillier, L., Durbin, R., Green, P., Shownkeen, R., Halloran, N., et al., 1992. A survey of expressed genes in Caenorhabditis elegans. Nat. Genet. 1, 114–123.

Weber, J.L., Myers, E.W., 1997. Human whole-genome shotgun sequencing. Genome Res. 7, 401–409.

Welsh, J., Chada, K., Dalal, S.S., Cheng, R., Ralph, D., McClelland, M., 1992. Arbitrarily primed PCR fingerprinting of RNA. Nucleic Acids Res. 20, 4965–4970.

White, O., Kerlavage, A.R., 1996. TDB: new databases for biological discovery. Methods Enzymol. 266, 27–40.

Wiemann, S., Voss, H., Schwager, C., Rupp, T., Stegemann, J., Zimmermann, J., Grothues, D., Sensen, C., Erfle, H., Hewitt, N., et al., 1993. Sequencing and analysis of 51.6 kilobases on the left arm of chromosome XI from Saccharomyces cerevisiae reveals 23 open reading frames including the FAS1 gene. Yeast 9, 1343–1348.

Wiemann, S., Stegemann, J., Grothues, D., Bosch, A., Estivill, X., Schwager, C., Zimmermann, J., Voss, H., Ansorge, W., 1995. Simultaneous on-line DNA sequencing on both strands with two fluorescent dyes. Anal. Biochem. 224, 117–121.

Wilcox, A.S., Khan, A.S., Hopkins, J.A., Sikela, J.M., 1991. Use of 3′ untranslated sequences of human cDNAs for rapid chromosome assignment and conversion to STSs: implications for an expression map of the genome. Nucleic Acids Res. 19, 1837–1843.

Wilson, R.K., Koop, B.F., Chen, C., Halloran, N., Sciammis, R., Hood, L., 1992. Nucleotide sequence analysis of 95 kb near the 3′ end of the murine T-cell receptor alpha/delta chain locus: strategy and methodology. Genomics 13, 1198–1208.

Wilson, R., Ainscough, R., Anderson, K., Baynes, C., Berks, M., Bonfield, J., Burton, J., Connell, M., Copsey, T., Cooper, J., et al., 1994. 2.2 Mb of contiguous nucleotide

sequence from chromosome III of C. elegans. Nature 368, 32–38.

Wolfsberg, T.G., Landsman, D., 1997. A comparison of expressed sequence tags (ESTs) to human genomic sequences. Nucleic Acids Res. 25, 1626–1632.

Yu, W., Andersson, B., Worley, K.C., Muzny, D.M., Ding, Y., Liu, W., Ricafrente, J.Y., Wentland, M.A., Lennon, G., Gibbs, R.A., 1997. Large-scale concatenation cDNA sequencing. Genome Res. 7, 353–358.

Zhang, L., Zhou, W., Velculescu, V.E., Kern, S.E., Hruban, R.H., Hamilton, S.R., Vogelstein, B., Kinzler, K.W., 1997. Gene expression profiles in normal and cancer cells. Science 276, 1268–1272.

Zimmermann, J., Voss, H., Schwager, C., Stegemann, J., Erfle, H., Stucky, K., Kristensen, T., Ansorge, W., 1990. A simplified protocol for fast plasmid DNA sequencing. Nucleic Acids Res. 18, 1067.