

2 The Human Genome Project

LAP CHEE TSUI
STEVE W. SCHERER

Toronto, Canada

- 1 Introduction 42
- 2 Chromosome Maps 42
 - 2.1 Genetic Maps 43
 - 2.2 Physical Maps 44
- 3 DNA Sequencing 45
 - 3.1 cDNA Sequencing 47
 - 3.2 Systematic Mapping and DNA Sequencing of Human Chromosomes 47
 - 3.3 Whole Genome Shotgun 48
- 4 Annotation of the Genome DNA Sequence 49
 - 4.1 Cataloguing the Genes 51
 - 4.2 Disease Gene Identification 53
 - 4.3 Human Genome Sequence Variation 55
- 5 Conclusion 55
- 6 References 56

1 Introduction

The Human Genome Project (HGP) is a worldwide research initiative with the goal of analyzing the complete sequence of human DNA to identify all of the genes. Further understanding of the structure and organization of genes will allow for a systematic analysis of their normal function and regulation in an organism. A comprehensive description of the human genome is thus the foundation of human biology and the essential prerequisite for an in-depth understanding of disease mechanisms. As such, information generated by the HGP will represent a source book for biomedical science in the 21st century. It will help scientists and clinicians alike to understand, diagnose, and eventually treat many of the 5,000 genetic diseases that afflict humankind, including the multifactorial diseases in which genetic predisposition plays an important role.

With the introduction of somatic cell technology, recombinant DNA, and polymorphic DNA markers, human geneticists began to dissect the human genome at the molecular level in the early 1980s. Due to the limited resources and knowledge base available, most efforts were concentrated only in scattered regions of chromosomes with known biological significance or disease relevance. Every project had to invest an enormous amount of effort in setting up the technology and recruiting the appropriate resources before any reasonable progress could be made. Also, it was realized that detailed characterization of chromosomes and isolation of disease genes were just the first steps required to approach the biological problems. It soon became apparent that a large-scale effort would be more efficient and cost-saving, if the ultimate goal was to identify all the molecular defects in diseases, as well as the structure and function of the genes in the human genome (COOK-DEEGAN, 1989; SINSHEIMER, 1989; DULBECCO, 1986). It was also clear that such a large-scale study would necessitate collaboration of research laboratories at the international level (CANTOR, 1990).

The international HGP was officially launched in the early 1990s. Its progress and general strategy can be described in three pha-

ses (outlined in Fig. 1). These include (1) the generation of chromosome maps, (2) large-scale DNA sequencing, and (3) annotating the DNA sequence (GUYER and COLLINS, 1995; COLLINS et al., 1998). Prior to the initiation of each of these phases as well as the HGP itself, significant advances in technology were necessary (Fig. 1). Moreover, due to these technological advances the international HGP has experienced changes in strategy. There were several genome-wide efforts to complete each of these stages, but to monitor the true progress it was often easiest to measure the level of completeness in chromosome units. Presently, almost all regions of chromosomes have high resolution maps, and a "working draft" DNA sequence of the human genome has been assembled. The DNA sequence of chromosomes 21 and 22 is complete (HATTORI et al., 2000; DUNHAM et al., 2000). The next five years of the HGP will involve annotating the DNA sequence. This will include completing the DNA sequence, characterizing all of the genes, identifying DNA sequence variations and mutations associated with disease, and collating all of the resulting data as a reference for future biomedical studies.

2 Chromosome Maps

There are two types of chromosome maps: physical maps and genetic maps. Physical mapping uses a variety of methods to assign genes and DNA markers to particular locations along a chromosome, so the actual distances between the genes (measured in nucleotide base pairs) are known (as discussed below, the finished DNA sequence is the highest resolution physical map, but to achieve an ordered sequence map lower resolution maps are first required). Genetic mapping describes the arrangement of genes based on the relationship of their linkage. DNA markers or probes can also be used in the construction of genetic maps, if they detect sequence changes (polymorphism) among different individuals. The tendency of two genes or DNA markers to segregate together through meiosis in family studies gives a description of genetic linkage, but

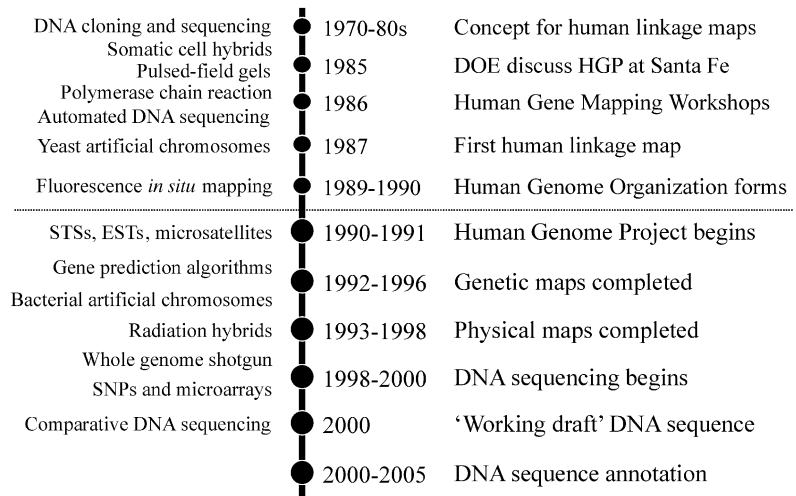


Fig. 1. Timeline of the Human Genome Project (HGP). A historical summary of the enabling technologies (on the left) and the achievements (on the right) leading up to and including the HGP are shown. The formal international HGP began around 1990. Other aspects of the HGP are the study of model organisms, analysis of genome variation, and the development of bioinformatics. Establishing training and public education programs, as well as the study of the ethical, legal, and social issues of genetics research in society are also important priorities.

not their physical location. The order of genes on a chromosome measured by genetic linkage is the same as the order in physical maps, but there is no constant scale factor that relates physical and genetic distance. The variation in scale occurs because recombination does not occur at equal frequencies for different intervals along a chromosome. Since most of the genes and DNA markers used in the construction of genetic maps exist as cloned and sequenced DNA fragments, they can also be readily placed on a physical map.

2.1 Genetic Maps

Due to the initial interest in disease gene cloning, most of the early efforts (pre-HGP) in generating genetic maps were focussed on the isolation of genetic markers that could be used in linkage analysis for mapping of disease loci. Throughout the course of the HGP, the choice of genetic markers evolved from restriction

fragment length polymorphisms (RFLPs) (KAN and DOZY, 1980; BOTSTEIN et al., 1980; WHITE et al., 1985; DONIS-KELLER et al., 1987), to variable number tandem repeats (VNTRs) (NAKAMURA et al., 1987), to microsatellites (length variation of simple di-, tri-, and tetra-nucleotide repeats) (WEBER and MAY, 1989; LITT and LUTY, 1989). By following the inheritance of marker alleles, linkage groups were constructed on each chromosome. The first example of a genome-wide effort was the establishment of the Centre d'Etude du Polymorphisme Humain (CEPH) in 1984 to produce reference maps with a set of well-defined pedigrees (DAUSSET, 1986). Since microsatellites are highly polymorphic (ideal for linkage analysis) and abundant in the genome, extensive efforts were devoted to the isolation of these markers and their use in the generation of genetic maps. Studies from Genethon and CEPH maps (WEISSENBACH et al., 1992; GYAPAY et al., 1994; DIB et al., 1996), the Collaborative (DONIS-KELLER et al., 1992) and Co-operative

Human Linkage Centers (BUETOW et al., 1994), and the Utah Marker Development Group (The Utah Marker Development Group, 1995), generated human genetic maps at the 1 centiMorgan (cM) resolution level. These genetic maps provided essential markers for constructing accurate physical maps of the chromosomes. Single nucleotide polymorphisms (SNPs) represent the most common polymorphism in the human genome occurring on average, once in every 1,000 base pairs (bp) (COLLINS et al., 1997). Their abundance and amenability for automated analysis by high-throughput technologies will allow genome-wide association studies to be conducted in projects aimed at identifying genes involved in multifactorial diseases. High-density SNP maps of human chromosomes are currently being constructed (KRUGLYAK, 1997; WANG et al., 1998).

2.2 Physical Maps

Early physical mapping studies were limited to the use of *in situ* hybridization of DNA probes to metaphase chromosome spreads and gel-blot hybridization analysis of DNA digested with various combinations of restriction enzymes. The introduction of pulsed field gel electrophoresis (PFGE) (SCHWARTZ and CANTOR, 1984) and fluorescence *in situ* hybridization (FISH) (TRASK et al., 1989; LAWRENCE, 1990; LICHTER et al., 1990; LAWRENCE et al., 1990; Pinkel et al. 1986) were major technology advances, but significant improvement of the mapping strategy only came after the availability of the yeast artificial chromosome (YAC) (BURKE et al., 1987) cloning system. YAC vectors are capable of propagating large fragments of human DNA insert 500,000 bp (500 kb) to 1 million bp (1 Mb) in size in yeast. Any genomic DNA fragments and cDNA (copies of messenger RNA) clones could be used to assemble overlapping clones (contigs) by hybridization screening. The introduction of sequence tagged sites (STSs), which are short DNA segments defined by their unique sequences, allowed the use of polymerase chain reaction (PCR) in contig assembly (known as anchor loci mapping or STS-content mapping; OLSON et al., 1989). STSs that marked genes

were called expressed sequence tags (ESTs) (ADAMS et al., 1991; OKUBO et al., 1992). DNA fingerprinting is another method of assembling overlapping cloned DNA molecules based on restriction fragment analysis or Southern blot hybridization patterns (KOHARA et al., 1987; COULSON et al., 1986; OLSON et al., 1986; BELLANNÉ-CHANTELOT, 1992).

Another powerful large-scale physical mapping technique is the use of radiation-hybrid (RH) mapping (COX et al., 1990). Human chromosome fragments were generated by X-ray irradiation of somatic cells and maintained in rodent cell backgrounds through the use of cell fusion techniques. Since only a portion of the human chromosomes is retained in each hybrid cell line, statistical methods could be used to assess physical linkage relationships between DNA markers and build linkage groups. With the inclusion of anchor loci (genes, genetic markers, STSs) in the mapping, it was possible to rapidly position unknown genes or DNA markers to specific chromosome regions resulting in the generation of useful RH maps of the human genome (WALTER et al., 1994).

Much of the early success of YAC-based contig mapping came from defined chromosome regions and some of the smaller chromosomes (GREEN and OLSON, 1990; CHUMAKOV et al., 1992; FOOTE et al., 1992). Considerable excitement was generated from the whole genome mapping work at CEPH-Genethon and subsequently the Whitehead Genome Center (COHEN et al., 1993; CHUMAKOV et al., 1995; HUDSON et al., 1995). The two groups generated YAC contig coverage initially estimated to be 75% and 95% complete for the entire genome, respectively. It became clear, however, that the level of coverage reported by both groups was probably overestimated. The overestimation was primarily caused by a paucity of DNA markers for some chromosomal regions and the high percentage (40–50%) of YAC clones in the libraries containing non-contiguous DNA sequences (chimeric clones) which were not readily detectable by the algorithms when used without additional analysis. In addition, there exist regions of human chromosomes not intrinsically amenable to cloning in yeast. Nevertheless, the data generated in both efforts was extremely valuable for dis-

ease gene identification studies and it provided an ordered scaffold of DNA markers which could be used in the DNA sequencing stage of the HGP (Fig. 2). The development of bacterial-based cloning vectors (BACs) (IOANNOU et al., 1994; SHIZUYA et al., 1992; see below) was important since the majority of DNA sequences that could not be cloned or mapped using YACs, could be analyzed using this system. Bacterial-based DNA molecules also provided the substrate for the clone-by-clone approach to sequence the human genome (Fig. 2).

Therefore, the genetic and physical mapping resources generated by the many different experimental strategies, from both chromosome-specific and whole-genome efforts, were complementary and essential for the success of the HGP. The establishment of practical, accurate, and detailed maps of ordered DNA markers along each chromosome provided the framework for DNA sequence maps (see Sect. 3). Moreover, since many of these DNA markers and clones were used in biological studies, they remain useful for phase III of the HGP; annotating the DNA sequence. The earliest examples of integrated physical and genetic maps were for the smallest chromosomes, 21 (CHUMAKOV et al., 1995) and 22 (COLLINS et al., 1995). Beginning in 1973 and throughout the HGP, single chromosome data, maps, and DNA sequences were collated at the International Gene Mapping Workshops and Single Chromosome Workshops often sponsored by

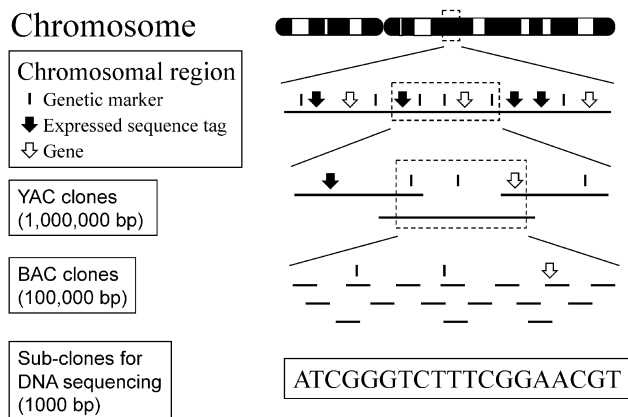
the Human Genome Organization (HUGO) (MCKUSICK, 1989). This data was submitted to relevant scientific databases such as the Genome Database and GenBank (see Tab. 1).

3 DNA Sequencing

The haploid human genome is estimated to contain approximately 3 billion bp of DNA sequence with the smallest chromosomes (21, 22, and Y) being about 40–50 Mb and the larger chromosomes (1 and 2) being 260 Mb; the average size chromosome is 130 Mb, approximately equal in size to the entire genome of the fruitfly *Drosophila melanogaster* (MYERS et al., 2000). If unwound and tied together, the strands of adenines, thymines, cytosines, and guanines (A,C,G,Ts) comprising human DNA, would stretch more than 5 feet but would be only 50 trillionths of an inch wide.

Due to the obvious enormity of the task at hand, it became clear that large-scale genomic DNA sequencing would be most economical when conducted in organized and automated facilities with substantial computerization and instrumentation (OLSON, 1995). There were 20 such centers involved in the international human sequencing consortium with the largest contributions coming from the Sanger Center in Cambridge, UK, Washington University in St. Louis, USA, and the Whitehead Institute in

Fig. 2. Systematic mapping and sequencing of human chromosomes using a clone-by-clone approach. DNA markers from a chromosomal region are positioned and simultaneously used to order large cloned DNA molecules (YACs and BACs). BACs are broken down in “sub-clones” for DNA sequencing. Obtaining partial sequence of the BACs represents “working draft” sequence. The DNA sequence is the highest resolution physical map.



Boston, USA. While a number of sequencing technologies were being developed, one of the original methodologies described in the 1970s by SANGER and colleagues (SANGER et al., 1977) provided the basic chemistry (fluorescence-based dideoxy sequencing) (HUNKAPILLER et al., 1991) performed in the HGP. The principles behind the Sanger method include using (1) an enzymatic procedure to synthesize DNA chains of varying lengths that terminate at either the A, T, C, or G nucleotides, and (2) separating the fragments on gels by electrophoresis to determine the identity and order of nucleotides based on the size of the fragment.

Due to limitations of DNA isolation and sequencing technologies, however, the ordered DNA molecules cloned into YAC contigs (which comprised the first-generation physical maps of the human genome) had to first be converted to smaller-size clones to prepare for genomic DNA sequence determination (see Fig. 2). Therefore, the sequencing protocols currently used are almost exclusively based on

using clones from existing bacterial-based cloning libraries (see below). If necessary, YAC clones could be subcloned directly in plasmid vectors for shotgun sequencing of regions, as was demonstrated in the *C. elegans* genome sequencing project (The *C. elegans* Sequencing Consortium, 1998). For the human genome project BAC cloning systems were primarily used. BACs are capable of carrying inserts in the 100–500 kb size range, with the libraries used for the HGP having inserts of about 150 kb.

Two general strategies were followed to yield a “working draft” DNA sequence of the human genome. These include using a systematic clone-by-clone approach (Fig. 2) and the whole-genome shotgun strategy (WEBER and MYERS, 1997; VENTER et al., 1998; Fig. 3). In addition to genomic DNA sequencing, many cDNA sequencing projects were initiated to capture information about protein coding genes.

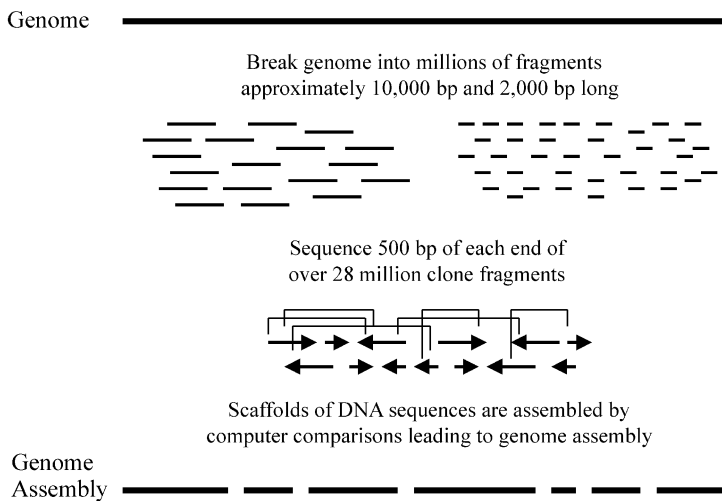


Fig. 3. Whole Genome Shotgun (WGS) DNA sequencing strategy of the human genome. The genome is broken into random DNA fragments that are cloned and sequenced from each end. Large scaffolds of DNA sequences can be assembled by identifying overlapping stretches of DNA sequence and these can be ordered and extended because both ends of the clone are known. This strategy bypasses the physical mapping stage used in the clone-by-clone approach (Fig. 2). However, for the human genome the assembled DNA sequences need to be positioned on chromosomes by comparison to DNA markers derived from well-characterized genetic and physical maps.

3.1 cDNA Sequencing

Prior to starting large-scale genomic sequencing projects, many initiatives were undertaken to capture (through DNA sequencing) those portions of the genome that are transcribed into genes. The reason for this is that only 5% of the DNA contained on a chromosome encodes for genes. These DNA sequences are called exons and the intervening genomic sequences are called introns (see Fig. 5, for example). By isolating messenger RNA (mRNA) from human tissues, the exon coding sequences were captured as cDNA allowing rapid isolation of putative gene sequences. The resulting cDNA clones were sequenced and ESTs for genes could be made. The first major public cDNA sequencing effort was initiated at the Institute for Genomic Research (TIGR) (ADAMS et al., 1995). Subsequently, the Integrated Molecular Analysis of Gene Expression (IMAGE) Consortium and the Merck-Washington University (WU) Initiative were formed (HILLIER et al., 1996). Since the introduction of the EST concept in 1991 the number of EST entries in public databases grew to over 2,800,000. These clones are derived from hundreds of different cDNA libraries constructed from most human tissues. Obviously, there is a substantial redundancy in these ESTs and they can be grouped into consensus groups or "Unigenes" based on sequence overlap. A consortium of investigators was formed and in one effort over 30,000 ESTs were mapped to defined chromosomal regions providing gene-based DNA markers for constructing YAC and BAC contigs (DELOUKAS et al., 1998). Despite these efforts, comprehensive gene identification, and in particular identifying cDNA covering the entire gene, almost always requires additional experimentation and verification. As part of the HGP, experiments for full-length gene identification and characterization through DNA sequencing have been undertaken the results of which will contribute substantially in annotating the genome sequence (STRAUSBERG et al., 1999).

3.2 Systematic Mapping and DNA Sequencing of Human Chromosomes

To provide the starting material for DNA sequencing of complex genomes, the standard approach relies on first building sequence-ready clone maps over regions ranging in size from hundreds of kilobases to entire chromosomes (see Fig. 2). As described in Sect. 2, the HGP BAC clone maps are assembled and ordered based on their DNA marker content as well as using fingerprinting techniques (GREGORY et al., 1997; MARRA et al., 1997) and by determining the DNA sequences of the ends of clones (VENTER et al., 1996). Additional mapping information on the chromosome location of the BACs along the chromosome could be determined using the clones directly as probes for hybridization analysis of chromosomes (see Fig. 5, for example). The BAC map of the human genome consists of >330,000 clones grouped into 1,743 contigs covering upwards of 95% of the genome.

For sequencing, 32,221 clones forming the minimal set covering the maximum region (a "tiling path") of each human chromosome was selected. Each clone then needs to be broken down into smaller subclones for sequencing. Initially, the plan was to put the fragments to be sequenced in order, followed by complete sequence determination of each fragment in a systematic manner so that the entire human DNA sequence of the BAC (and, therefore, the corresponding region on the chromosome) was known. This method produces highly accurate sequence with few gaps. However, the upfront process of building the sequence-ready maps, subclone library construction, and directed gap filling is costly, time consuming, and, therefore, often rate-limiting.

To accelerate progress, in 1998 a major deviation from this plan was the decision to collect only partial data from each DNA fragment, hence, a "working" or "rough" draft (GUYER and COLLINS, 1995) (the change of strategy was partly due to the launching of a privately funded company, Celera, who decided to determine the DNA sequence of small pieces of human DNA totally at random using the whole genome shotgun approach de-

scribed below). Working draft DNA sequence usually covers 95% of the BAC (maintaining 99% accuracy), but it is divided into 10–100 largely unordered segments. Additional sequencing is required to generate the finished DNA sequence such that there are no gaps or ambiguities and the final product is greater than 99.99% accurate. At the time of writing this chapter, close to 90% of the euchromatic human genome sequence was determined (approximately 75% of this was working draft and 25% in finished form). Results from the finished DNA sequence of chromosomes 21 and 22 are likely to be instructive for what a finished product of the human genome might look like. Chromosome 22q (an acrocentric chromosome) was determined to contain 33.5 Mb represented in 10 contigs. Although usually small, the gaps in the DNA sequence map could not be closed even after exhaustive attempts using multiple cloning systems (DUNHAM et al., 2000). For chromosome 21, 33.8 Mb of DNA sequence could be assembled and represented in three contigs with only a few detectable small gaps (HATTORI et al., 2000).

3.3 Whole Genome Shotgun

The whole genome shotgun strategy (WGS) involves shearing all of the DNA of an organism into segments of a defined length which are cloned into a plasmid vector for DNA sequencing (WEBER and MYERS, 1997; VENTER et al., 1998). Sufficient DNA sequencing is performed so that each nucleotide of DNA in the genome is covered numerous times in fragments of about 500 bp. After sequencing, the fragments are assembled to reconstruct the complete genome. WGS was used by scientists at The Institute for Genomic Research to generate the first complete sequence of a self-replicating organism called *Haemophilus influenzae* (FLEISCHMANN et al., 1995) as well as many other prokaryotic organisms. The advantage of WGS is that the upfront steps of constructing a physical map are not completed. For organisms with much larger and more complex genomes, such as *Drosophila melanogaster* and human, assemblies of sequences are expected to be complicated by the presence of a vast number of repetitive elements (approx-

mately 50% of human DNA is repeats). Notwithstanding, VENTER and colleagues at Celera Corporation initiated a WGS project of the *Drosophila* genome and in doing so, 3.2 million sequence reads were completed (giving a 12.8X coverage of the 120 Mb genome). Based on this data 115 Mb of DNA sequence could be assembled and although it was quite comprehensive, the genome was still divided by 1,630 gaps (MYERS et al., 2000). Closure of the gaps is being completed using the scaffold of BACs generated from physical mapping projects.

Following the experience gained from the *Drosophila melanogaster* project, Celera Corporation planned a whole-genome assembly of human DNA (VENTER et al., 1998). The aim of the project was to produce highly accurate, ordered sequence spanning more than 99.9% of the human genome. Based on the size of the human genome and the results from the *Drosophila* experiment it was predicted that over 70 million sequencing reactions would need to be completed. This would be divided into sequencing both ends of 30 million 2 kb clones, 5 million 10 kb clones, and 300,000 150 kb clones. The alignment of the resulting sequence assemblies along the chromosomes would be accomplished using the large number of DNA markers and physical maps generated by the ongoing HGP (see Sect. 2). Since efforts were escalated by the publicly funded HGP using the clone-by-clone “working draft” approach, Celera could also easily integrate this data into their assemblies, thereby greatly reducing the amount of sequencing required. Moreover, the sequencing of the ends of the 150 kb clones (which were BACs) was completed by publicly funded efforts. In the end, to assemble sequence contigs, Celera completed approximately 28 million reads and merged this data with DNA sequence in public databases (Fig. 4). The final sequence contigs were ordered using the DNA markers and maps from the HGP (see Sect. 2).

The completeness and accuracy of the first draft of the human genome sequence will be tested by many types of experimentation over the next decade. Based on the results from chromosome 21 and 22 it is expected that there will be chromosome regions that will not be represented, since some DNA cannot be cloned

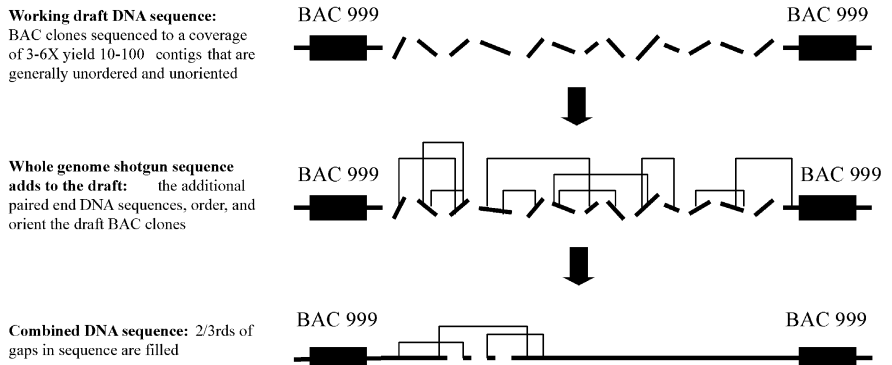


Fig. 4. Working draft DNA sequence combined with the whole genome shotgun (WGS) strategy. The strategy of the publicly funded HGP using mapped clones (Fig. 2) occurs in two phases, (1) the “shotgun” stage which involves determining most of the sequence from a clone that is assembled into a product (“working draft”) that contains gaps and ambiguities and (2) the “finishing” stage which involves additional directed sequencing for gap closure and resolution of ambiguities. DNA sequence data from the WGS strategy (see Fig. 3) can expedite the “finishing” stage since it can often extend contigs, fill gaps, and resolve orders of sequence scaffolds. Some WGS sequences could also reside in regions of the human difficult to clone using the existing vectors.

using currently available vectors. Also, incorrect assemblies of sequence will occur due to the presence of repeats and duplications. It will be interesting to determine if data derived from the WGS approach can close some of the gaps that could not be completed using directed chromosome walking and sequencing. An example of the genomic representation of working draft DNA sequence alone compared to the equivalent sequence combined with Celera WGS data for the q35 region of human chromosome 7 is shown in Fig. 5.

4 Annotation of the Genome DNA Sequence

A complete understanding of the biology and function of the genome will be the ultimate goal of the HGP. The DNA sequence will undergo continual refinement as new sequen-

ces and new types of biological data are added. With proper annotation of the DNA sequence, a full description of all the genes and other important biological information stored within the DNA will be known. Given this information it will be possible to investigate the roles of all of the gene products, how they are controlled and interact, and their possible involvement in disease. The process of annotating the human DNA sequence will take several forms including (1) cataloguing all of the genes, (2) identifying the genes and DNA sequence variations that either directly cause or are associated with disease, (3) studying genetic variation, and (4) establishing integrated and curated databases containing all DNA sequence annotation. A full description of the databases is beyond the context of this chapter. We have summarized and provided links to relevant www sites where initiatives are ongoing to establish integrated databases (Tab. 1). A vital next step will be to organize experts in different fields of genomics for proper curation of this information.

